STUDIES IN SOFTWARE RELIABILITY GROWTH MODELS AND PROPORTIONAL HAZARD MODELS

THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MATHEMATICS

TO THE BUNDELKHAND UNIVERSITY JHANSI

By

HARENDRA SINGH YADAV
DEPTT. OF MATHEMATICS & STATISTICS

BUNDELKHAND UNIVERSITY, JHANSI

INDIA

DECEMBER 1994

DECLARATION

I declare that this research work has been carried out by me and no part of this thesis formed the basis for the award of any Degree, Diploma, Associateship or any other Similar title to me.

31/12/94

HARENDRA SINGH YADAV
Department of Mathematics & Statistics,
Bundelkhand University
Jhansi (U.P.)

DECEMBER-1994

CERTIFICATE

Certified that the work of the thesis entitled "STUDIES OF SOFTWARE RELIABILITY GROWTH MODELS AND PROPORTIONAL HAZARD MODELS" is submitted in Bundelkhand University, Jhansi by **Mr. Harendra Singh Yadav**, for the award of the degree of Doctor of Philosophy is based on his research work carried out under my supervision and guidance.

This work either in part or in full has not been submitted to any university or institution for the award of any degree.

VKSehgal 31/12/94

DATED : 31-12-94

Dr. V.K. SEHGAL
SUPERVISOR
READER
DEPTT. OF MATHEMATICS & STATISTICS
BUNDELKHAND UNIVERSITY
JHANSI (U.P.)

ACKNOWLEDGEMENTS

I have great pleasure in expressing my deepest sense of gratitude and indebtedness to Dr. V. K. Sehgal, Reader in the Department of Mathematics and Statistics, Bundelkhand University, Jhansi (U.P.). I feel my self fortunate to get his unstinted cooperation and guidance through the entire course of my research work. His kind disposition favour and zeal for my work has been a source of constant inspiration to me.

I am grateful to Professor S. Biswas, Deptt. of Statistics, University of Delhi and Dr. P.K. Kapur, Reader, Department of Operational Research, University of Delhi for their encouragement and help during the course my research work.

I am also thankful to Dr. P.N. Shrivastava, Reader and Head, Deptt. of Mathematics & Statistics, Bundelkhand University, Jhansi for his support and useful suggestions.

31/12/94

DATE : 31.12.94

HARENDRA SINGH
YADAV

CONTENTS

CHAPTER 1
INTRODUCTION & BASIC CONCEPTS

1.0 : INTRODUCTION :

This chapter of the thesis is devoted to Basic Concepts of Reliability, Software Reliability, Baysian Inference and proportional hazard models. Reliability is an important consideration in the planning design and operation of the systems. The reliability theory is concerned with random occurrence of undesirable events or failures during the life time of a physical system or biological system. Reliability is an inherent attribute of a system just as the system capacity or power rating. The concept of reliability has been known for a number of years but has got greater significance and importance during the past decade particularly due to the impact of automation, development of complex missile and space programmes. In recent years the concept of reliability has been formulated as the science of predicting, estimating or optimizing the probability of survival, the mean life or more generally the life distribution of components or systems. As a result of advancement of science today man has to his credit so many sophisticated systems which are fully designed by his hands and brain. e.g. Television System, Computer System, Electric

Power Supply System etc are some man made system. Generally when we perform the life testing experiment with man made systems, we call it "Reliability Analysis" while on the other hand when we deal with god made systems like Human Body System, Weather Changing System, Solar Energy System etc, we name it "Survival Analysis". Reliability and Survival are interchange terms Reliability function and Survival function are represented by R(t) and S(t) respectively. Hoxford (1960) defined the reliability through the concept of dependability as the probability that the system will be able to operate when needful, while according to Bazovsky (1961) the reliability is a yardstick of the capability of an equipment to operate without failure when put into service. Reliability in a simplest form means the probability that a failure may not occur in a given time interval. Reliability of a component or a system is the probability that the components performs its intended function adequately for a specified period of time under the stated operating conditions or environment. In other words, Reliability Engineering is a branch of Science deals with the life testing experiment with engineering systems like, Radio, Television System, Electric Power Supply Systems and Computer System etc.

1.1 : RELIABILITY FUNCTION :

If T is a random variable, denotes the time to failure of component then the probability that it will not fail in a given environment before time t and thus reliability function can be written as

$$R(t) = P[T > t] = 1 - P[T \leq t]$$
$$= 1 - \bar{R}(t)$$
$$= 1 - F(t) = \bar{F}\ (t) \qquad \ldots\ldots(1.1)$$

where F(t) is the cumulative distribution function (c.d.f.) of T called unreliability $\bar{R}(t)$ of the component so that

$$R(t) + \bar{R}(t) = 1$$

Thus the reliability is a function of time and depends on environmental conditions which may or may not vary with time. Since the reliability of a unit is a probability, its numerical value always lies between 0 and 1.

$$\lim_{t \to 0} R(t) = 1 \qquad \text{and} \qquad \lim_{t \to \infty} R(t) = 0$$

1.2 : FAILURE TIME DISTRIBUTION :

Let T be a non negative random variable representing the failure time of an individual form a homogeneous population. The probability distribution of T can be specified in many ways, three of which are particularly useful in survival applications The Survivor Function, The Probability Density Function and The Hazard Function. Interrelations between these three representations

1.1 : RELIABILITY FUNCTION :

If T is a random variable, denotes the time to failure of component then the probability that it will not fail in a given environment before time t and thus reliability function can be written as

$$R(t) = P[T > t] = 1 - P[T \leq t]$$

$$= 1 - \bar{R}(t)$$

$$= 1 - F(t) = \bar{F}\ (t) \qquad \text{......(1.1)}$$

where F(t) is the cumulative distribution function (c.d.f.) of T called unreliability $\bar{R}(t)$ of the component so that

$$R(t) + \bar{R}(t) = 1$$

Thus the reliability is a function of time and depends on environmental conditions which may or may not vary with time. Since the reliability of a unit is a probability, its numerical value always lies between 0 and 1.

$$\lim_{t \to 0} R(t) = 1 \qquad \text{and} \qquad \lim_{t \to \infty} R(t) = 0$$

1.2 : FAILURE TIME DISTRIBUTION :

Let T be a non negative random variable representing the failure time of an individual form a homogeneous population. The probability distribution of T can be specified in many ways, three of which are particularly useful in survival applications The Survivor Function, The Probability Density Function and The Hazard Function. Interrelations between these three representations are given below for both discrete and continuous distributions. The survivor function is defined for both discrete and continuous distribution as the probability that T is at least as great as a value t, that is

$$S(t) = P(T \geq t), \quad 0<t<\infty \qquad \text{......(1.2)}$$

1.2.1 CASE I - ABSOLUTELY CONTINUOUS : The probability density function (p.d.f.) of T is

$$\begin{aligned} f(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \\ &= \frac{-dS(t)}{dt} \qquad \text{......(1.3)} \end{aligned}$$

Conversely, the Survivor Function S(t) given as

$$S(t) = \int_t^\infty f(s)\, ds$$

and

$$f(t) \geq 0 \text{ with } \int_0^\infty f(t)\, dt = 1$$

The range of T is $(0,\infty)$ and this should be understood as the domain of definition for function of t.

The hazard function specifies the instantaneous rate of failure at (T=t) conditional upon survival to time t and is defined as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \\ &= \frac{f(t)}{S(t)} \qquad \text{......(1.4)} \end{aligned}$$

where h(t) specifies the distribution of T, since from (1.4)

$$h(t) = -\frac{d}{dt} \log S(t)$$

so that on integrating and using S(0) = 1, we get,

$$S(t) = \exp\left[- \int_o^t h(u)\ du \right] \qquad \ldots\ldots(1.5)$$

The p.d.f. of T can be written

$$f(t) = h(t) \exp \left[- \int_o^t h(u)\ du \right] \qquad \ldots\ldots(1.6)$$

From equation (1.5), we can say that h(t) is a non-negative function with

$$\int_o^s h(u)\ du < \infty$$

$$\text{for some } s > 0, \int_o^\infty h(u)\ du = \infty$$

The expected residual life at time t is given by

$$r(t) = E[(T - t)|T \geq t];\ 0 \leq t < \infty \qquad \ldots\ldots(1.7)$$

which uniquely determines a continuous survival distribution with finite mean, since

$$r(t) = \int_t^\infty \frac{(u-t)\ f(u)}{S(t)}\ du$$

$$= \int_t^\infty \frac{S(u)}{S(t)}\ du \qquad \ldots\ldots(1.8)$$

On integrating by parts, we have

$$\frac{1}{r(t)} = \frac{d}{dt} \log\int_t^\infty S(u)\ du \qquad \ldots\ldots(1.9)$$

Substituting t=0 in (1.8), we get

$$r(0) = \int_o^\infty S(u)\ du$$

and

$$\int_o^t \frac{du}{r(u)} = -\log \int_t^\infty S(u)\ du + \log r(0)$$

which leads finally to

$$S(t) = \frac{r(0)}{r(t)} \exp\left[- \int_0^t \frac{du}{r(u)}\right] \qquad \ldots\ldots(1.10)$$

1.2.2 CASE-II WHEN T IS DISCRETE :

If T is a discrete taking values $x_1 < x_2 < \ldots\ldots$ with associated probability function

$$p(x_i) = P\ (T = x_i); \quad (i = 1,2,3,\ldots\ldots)$$

then the survivor function is

$$S(t) = \sum_{j|x \geq t} p(x_j) \qquad \ldots\ldots(1.11)$$

$$= \sum p(x_j)\ H(x_j - t) \qquad \ldots\ldots(1.12)$$

where H(x) is Heaviside function

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

The hazard at x_j is defined as the conditional probability of failure at x_j.

$$h_j = P(T = x_j | T \geq x_j)$$

$$= \frac{P(x_j)}{S(x_j)}; \quad j = 1,2,3,\ldots\ldots \qquad \ldots\ldots(1.13)$$

Corresponding to (1.5) & (1.6), the survivor function and the probability function are given by

$$S(t) = \prod_{j|x_j<t} (1 - h_j) \qquad \ldots\ldots(1.14)$$

and
$$p(x_j) = h_j \prod_{1}^{j-1} (1 - h_i) \qquad \ldots\ldots(1.15)$$

1.2.3 ESTIMATION OF THE SURVIVOR FUNCTION :

Let $\overline{F}_n(x)$ = be the sample cumulative distribution function $= \frac{\text{number of sample values} \leq x}{n}$

A plot of $\overline{F}_n(x)$ versus x visually represent the sample and provides full information on the percentile points, the dispersion and the general features of the sample distribution, it is an indispensable aid in studying the distribution shape of the population from which the sample arose, in fact the sample distribution function can serve as a basic tool in constructing formal tests of goodness of fit of the data to hypothesized probability models.

In the analysis of survival data it is very often useful to summarise the survival experience of particular groups of patients in terms of the sample c.d.f. or more usually in terms of the sample survivor function. If an uncensored sample of distinct failure times is observed from a homogeneous population, the sample survivor function is a step function decreasing by n^{-1} immediately following each observed failure time.

Let $t_1 < t_2 \ldots\ldots < t_k$ represent the observed failure time in a sample of size n_o from a homogeneous population with the survivor function $S(t)$, suppose that d_j items fail at $t_j (j=1,2,..k)$ and in items are censored in the interval (t_j, t_{j+1}) at times $t_{j1}, \ldots t_{jm_j}$ $(j = 0,1,\ldots,k)$ where $t_o = 0$ and $t_{k+1} = \infty$, Let $n_j = (m_j + d_j) + \ldots\ldots + (m_k + d_k)$ be the number of items at risk at a time just prior to t_j. The probability of failure at t_j is

$$S(t_j) - S(t_j + 0)$$

where $S(t_j + 0) = \lim_{x \to 0^+} S(t_j + x)$; $(j = 1,2,...,k)$

we assume that the contribution to the likelihood of a survival time censored at t_{jl} is

$$P\ (T > t_{jl}) = S(t_{jl} + 0)$$

In effect, we are assuming that the observed censoring time t_{jl} conditions under that the unobserved failure time is greater than t_{jl}.

If censoring time were fixed in advance for each time thus we obtain

$$L = \prod_{j=0}^{k} \left\{ \left[S(t_j) - S(t_j + 0) \right]^{d_j} \prod_{l=1}^{m_j} S(t_{jl} + 0) \right\}$$

which is a likelihood function on the space of survivor function S(t) for a given data. The maximum likelihood estimate is the survivor function $\hat{S}(t)$ that maximizes L. This definition of the maximum likelihood estimate is a generalization of the usual concept used in parametric models. There are dangers associated with maximising likelihood of many para since such techniques may lead to inefficient or inconsistent estimates. The results of such maximization require some investigation to assure they are reasonable.

Clearly $\hat{S}(t)$ is discontinues at the observed failure time since otherwise, $L = 0$, further subject to $t_{jl} \geq t_j$ $S(t_j + 0)$ is maximized by taking

$$S(t_{jl} + 0) = S(t_j + 0); \quad (j = 1,2,...,k;\ l = 1,2,...,m_j)$$

and $S(t_{ol}) = 1$, $(l = 1,\ldots,m_o)$ the function $\hat{S}(t)$ is then a discrete survivor function with Hazard components $\hat{h}_1\ldots\ldots\hat{h}_k$ at $(t_1\ldots t_k)$ respectively. Thus,

$$\hat{S}(t_j) = \prod_{l=1}^{j-1} (1 - \hat{h}_l) \qquad \ldots\ldots(1.16)$$

$$\hat{S}(t_j+0) = \prod_{l=1}^{j}(1-\hat{h}_l) \qquad \ldots\ldots(1.17)$$

where the $\hat{h}$ are chosen to maximize the function

$$L = \prod_{j=1}^{k} \left\{\left[\prod_{l=1}^{j-1}(1 - \hat{h}_l)^{d_j}\right]\left[\prod_{l=1}^{j} (1 - \hat{h}_l)^{m_j}\right]\right\}$$

$$= \prod_{j=1}^{k} \hat{h}_j^{d_j} (1 - \hat{h}_j)^{n_j-d_j} \qquad \ldots\ldots(1.18)$$

obtained by substitution of (1.16) and (1.17) in L.

$$\hat{h}_j = \frac{d_j}{n_j};\ (j = 1,2,\ldots\ldots,k)$$

and the product limit estimate of the survivor function is

$$\hat{S}(t) = \prod_{j\mid t_j<t} \left(\frac{n_j-d_j}{n_j}\right) \qquad \ldots\ldots(1.19)$$

The estimate $\hat{F}(t)$ is the direct generalisation of the sample survivor function for censored data. It was derived by Kaplan & Meier (1958) and is known as Kaplan Meier Estimate. The induced expression for the asymptotic variance of $\hat{S}(t)$ is then

$$\hat{Var}\ (\hat{S}(t)) = \hat{S}^2(t) \sum_{j\mid t_j<t} \frac{d_j}{n_j(n_j-d_j)} \qquad \ldots\ldots(1.20)$$

The expression (1.20) is known as Greenwood's formula (Greenwood (1926)) was first derived as the asymptotic variance of the classical life table estimator.

1.3 : PARAMETRIC FAILURE TIME MODELS :

The main interest of the section to consider relationship between failure time and explanatory variables. Therefore, we should consider failure time distribution for homogeneous population. Weibull and Exponential distributions are more often used parametric models on failure time data. These distributions admit closed form expressions for tail area probabilities. Log-normal and gamma distributions are still frequently applied to failure time data but less convenient computationally.

Let $T > 0$ is a random variable representing failure time and t represents a typical point in its range. We use

$$Y = \log T$$

to represent the log-failure time shape comparisons among the parametric models are often simpler in terms of Y than T.

1.3.1 THE EXPONENTIAL DISTRIBUTION :

The Exponential Distribution with one parameter is obtained if we consider the hazard function to be a constant.

$$h(t) = \theta \ ; \ \theta > 0 \text{ over the range of } T.$$

The instantaneous failure rate is independent of t. As we know that Exponential distribution has memoryless property.

Therefore, the conditional chance of failure in a time interval of specified length is the same regardless of how long the individual has been on trial. The survivor function of T is given by

$$S(t) = e^{-\int_{o}^{t} \lambda(u)\, du} = e^{-\int_{o}^{t} \theta\, du} = e^{-\theta t}$$

and its probability density function is

$$f(t;\theta) = \frac{-dS(t)}{dt} = \theta e^{-\theta t}; \quad t \geq 0 \qquad \text{......(1.21)}$$

The p.d.f. of $Y = \log_e T$ is obtained by means of the transformation

$$y = \log_e t$$

$$\Rightarrow \quad e^y = t$$

$$\Rightarrow \left|\frac{dt}{dy}\right| = e^y$$

Therefore, $g_y(y) = f(t) \left|\frac{dt}{dy}\right|$

$$= \theta e^{-\theta t}\, e^y = e^{y-\alpha-e^{(y-\alpha)}} \qquad \text{......(1.22)}$$

where $\qquad \alpha = -\log \theta \Rightarrow \theta = e^{-\alpha}$

Let us make the transformation $Y = \alpha + W$. The p.d.f. of W is given by

$$g_w(w) = e^{w-e^w}\ ; \ -\infty < w < \infty \qquad \text{......(1.23)}$$

which is an extreme value distribution. The exponential distribution arises, also as the limiting form of the distribution of minimum of samples from some densities with range on (0,a) for some a≤∞. This sometimes can be taken as theoretical justification for its use in survival studies in which a complex mechanism fails when any one of its many components fails.

1.3.2 THE WEIBULL DISTRIBUTION :

The Weibull Distribution with two parameters θ_1 and θ_2 is a generalisation of the exponential distribution with hazard function is given by

$$\lambda(t) = \theta_1\theta_2 \ (\theta_1 t)^{\theta_2 - 1} \text{ for } \theta_1, \theta_2 > 0$$

$\lambda(t)$ is a monotone decreasing for $\theta_2 < 1$ and increasing for $\theta_2 > 1$ and constant exponential hazard if $\theta_2 = 1$

The p.d.f. of T is given as

$$f(t) = \theta_1\theta_2 \ (\theta_1 t)^{\theta_2 - 1} \ e^{-(\theta_1 t)^{\theta_2}}; \ 0 \le t < \infty \qquad \text{......(1.24)}$$

and the survivor function is

$$S(t) = e^{-(\theta_1 t)^{\theta_2}} \qquad \text{......(1.25)}$$

$$\log(-\log S(t)) = \theta_2(\log t + \log \theta_1)$$

The plot of log $(-\log \hat{S}(t))$ versus log t gives empirical check for the Weibull Distribution where $\hat{S}(t)$ is a sample estimate of survivor function. The plot should give approximately a straight line, the slope of which provides a rough estimate of θ_2 and the log t intercept an estimate of log θ_1.

Therefore, the p.d.f. of $Y = \log_e T$ is

$$g_y(y) = \frac{1}{\sigma} \ e^{-\left[\frac{y-\alpha}{\sigma} - e^{(y-\alpha)/\sigma}\right]} ; \quad -\infty < y < \infty \qquad \text{......(1.26)}$$

where $\sigma = \dfrac{1}{\theta_2}$ and $\alpha = -\log \theta_1$

More simply, we can write

$$Y = \alpha + \sigma W$$

where W has extreme value p.d.f.

The shape of the density for Y is fixed because θ_1 and θ_2 affect only the location and the scaling of the distribution. The Weibull Distribution can also be developed as the limiting distribution of minimum of a sample from a continuous distribution with range on [0,u] for some u(0≤u<∞).

1.3.3 THE LOG-NORMAL DISTRIBUTION :

The density function of a log-normal variate T is given by

$$f(t) = \frac{1}{\sqrt{2\pi}} \frac{\theta_2}{t} e^{-\frac{\theta_2^2 (\log (\theta_1 t))^2}{2}} \qquad \ldots\ldots(1.27)$$

If we make the transformation $Y = \log_e T$

i.e. $y = \log_e t \Rightarrow e^y = t$

$$\frac{dt}{dy} = e^y$$

Then the p.d.f. of Y is given as

$$g(y) = \frac{1}{\sqrt{2\pi}} \frac{\theta_2}{t} e^{-\left[y+\theta_2^2 (\log \theta_1 + y)^2\right]/2}$$

$$g(y) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left[\frac{(y-\alpha)^2}{\sigma^2}\right]} \qquad \ldots\ldots(1.28)$$

Further, if we put $Y = \alpha + \sigma W$, where W is a standard normal variate with density function

$$\phi(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} \qquad \text{.......(1.29)}$$

The survivor function is given by

$$S(t) = \int_t^\infty f(u)\,du$$

$$= \int_{\log t}^\infty \frac{1}{\sqrt{2\pi}}\, \theta_2\, u^{-1} e^{-\frac{\theta_2^2(\log \theta_1 u)^2}{2}}\, du$$

put log u = y

$$\frac{1}{u}\,du = dy$$

$$= \frac{\theta_2}{\sqrt{2\pi}} \int_{\log t}^\infty e^{-\theta_2^2 \frac{(y + \log \theta_1)^2}{2}}\, dy$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{\log t}^\infty e^{-\frac{1}{2}\frac{(y-\alpha)^2}{\sigma^2}}\, dy$$

where $\alpha = -\log \theta_1$

$$\sigma^{-1} = \theta_2$$

$$\frac{y-\alpha}{\sigma} = w$$

$$dy = \sigma\, dw$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\left(\frac{\log t-\alpha}{\sigma}\right)}^\infty e^{-w^2/2}\, dw$$

$$= \int_{\theta_2(\log \theta_1 t)}^\infty \frac{1}{\sqrt{2\pi}} e^{-w^2/2}\, dw = 1 - \int_{-\infty}^{\theta_2 \log \theta_1 t} \frac{e^{-w^2/2}}{\sqrt{2\pi}}\, dw$$

$$= 1 - \phi(\theta_2 \log \theta_1 t) \qquad \text{.......(1.30)}$$

where $\phi(w) = \int_{-\infty}^{w} \phi(u)\, du$

The hazard function is $h(t) = \frac{f(t)}{S(t)}$

The hazard function has value 0 at t=0 increases to maximum and then decreases, approaching zero as t becomes large.

The log-normal model is particularly simple to apply if there is no censoring but with censoring the computations become difficult.

1.3.4 THE GAMMA DISTRIBUTION :

The Gamma Distribution is a two parameter generalisation of exponential distribution with density function

$$f(t;\theta_1,\theta_2) = \frac{\theta_1^{\theta_2}(t)^{\theta_2-1} e^{-\theta_1 t}}{\Gamma(\theta_2)}; \ \theta_1,\theta_2>0, 0\leq t<\infty \quad ....(1.31)$$

If we make the transformation Y = log T

i.e. $y = \log t, \Rightarrow t = e^y \qquad \frac{dt}{dy} = e^y$

Then the p.d.f. of Y = log T is given by

$$g(y) = \frac{\theta_1^{\theta_2}(e^y)^{\theta_2-1} e^{-\theta_1 e^y}}{\Gamma(\theta_2)} . e^y$$

$$= \frac{\theta_1^{\theta_2} e^{(\theta_2 y-\theta_1 e^y)}}{\Gamma(\theta_2)}$$

$$= \frac{e^{(\theta_2(y-\alpha)e^{y-\alpha})}}{\Gamma(\theta_2)}$$

where $\alpha = -\log \theta_1 \Rightarrow \theta_1 = e^{-\alpha}$

Further if $Y = \alpha + W$, the density function of W is

$$g(w) = \frac{e^{(\theta_2 w-e^v)}}{\Gamma(\theta_2)} \quad ......(1.32)$$

The error quantity W has a negatively skewed distribution with skewness decreasing with increasing θ_2. At the exponential model $\theta_2=1$. The survivor function of Gamma Distribution is

$$S(t) = P[T > t] = \int_t^\infty f(u)\, du$$

$$= 1 - \int_0^t f(u)\, du$$

$$= 1 - \frac{\theta_1^{\theta_2}}{\Gamma(\theta_2)} \int_0^t e^{-\theta_1 u}\, u^{\theta_2 -1}\, dt$$

$$= 1 - \frac{\theta_1^{\theta_2}}{\Gamma(\theta_2)} \int_0^{\theta_1 t} e^{-v}\, v^{\theta_2 -1}\, dv \qquad \text{Put } \theta_1 u = v,\ \theta_1 du = dv$$

$$= 1 -1\ I_{\theta_2}(\theta_1 t) \qquad \ldots\ldots(1.33)$$

where $I_{\theta_2}(S)$ is the incomplete Gamma Integral

$$I_{\theta_2}(S) = \frac{1}{\Gamma(\theta_2)} \int_0^S x^{\theta_2 -1}\, e^{-x}\, dx \qquad \ldots\ldots(1.34)$$

and the hazard function h(t) is given by

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta_1 (\theta_1 t)^{\theta_2 -1}\, e^{-\theta_1 t}\, [\Gamma(\theta_2)]^{-1}}{1-I_{\theta_2}(\theta_1 t)} \qquad ..(1.35)$$

The hazard function is monotone increasing from 0 if $\theta_2>1$ monotone decreasing from ∞ if $\theta_2<1$ and in either case approaches θ_1 as t becomes large.

If $\theta_2=1$, the Gamma Distribution reduces to the exponential distribution, with integer θ_2, the gamma distribution is also known as a special Erlangian Distribution. The Gamma distribution with integer θ_2 can also be derived as the waiting time to the θ_2th emission from a Poisson source with intensity parameter θ_1. The sum of θ_2 independent exponential variates with failure rate θ_1 has also the Gamma Distribution with parameters θ_1 and θ_2.

1.3.5 LOG-LOGISTIC DISTRIBUTION :

Consider the model $Y = \log T = \alpha + \sigma W$

where $\alpha = -\log \theta_1 \sigma = \theta_2^{-1}$

We can construct different failure time models by selecting different distribution for the error variable W. One such is the log-logistic distribution for T if W has the logistic density

$$g(W) = \frac{e^{W}}{(1+e^{W})^2} \qquad ....(1.36)$$

This is a symmetric density with mean and variance given by

$$E(W) = 0 \qquad V(W) = \frac{\pi^2}{3}$$

The p.d.f. of T is then

$$f(t;\theta_1,\theta_2)=\theta_1\theta_2(\theta_1 t)^{\theta_2 -1}[1+(\theta_1 t)^{\theta_2}]^{-2};\ 0\le t<\infty \ ......(1.37)$$

where $\theta_1 = e^{-\alpha}$ and $\theta_2 = \sigma^{-1}$

The survivor and hazard functions are given by

$$S(t) = \frac{1}{1+(\theta_1 t)^{\theta_2}} \qquad ......(1.38)$$

$$h(t) = \frac{\theta_1\theta_2(\theta_1 t)^{\theta_2 -1}}{1+(\theta_1 t)^{\theta_2}} \qquad ......(1.39)$$

The distribution is more useful for handling censored data than the log-normal distribution while providing a good approximation to it except in the extreme tails. The hazard function is identical to the Weibull hazard aside from the denominator factor $1+(\theta_1 t)^{\theta_2}$. It is monotone decreasing from ∞ if $\theta_2<1$ and is monotone increasing from θ_1 if $\theta_2=1$. If $\theta_2>1$, the hazard function resembles the log-normal hazard in that it increases from zero to a maximum at $t = \frac{(\theta_2-1)^{1/\theta_2}}{\theta_1}$ and decreases toward zero thereafter.

1.3.6 GENERALISED GAMMA DISTRIBUTION :

The p.d.f. of three parameter generalised distribution is given by

$$f(t;\theta_1,\theta_2,\theta_3) = \frac{\theta_1\theta_2(\theta_1 t)^{\theta_2\theta_3-1}\ e^{-(\theta_1 t)^{\theta_2}}}{\Gamma(\theta_3)};\ t>0 \ .....(1.40)$$

where $\theta_1 = e^{-\alpha}$ and $\theta_2 = \sigma^{-1}$

This model was introduced by Stacy (1962).

Special Cases :

(i) When $\theta_2=1=\theta_3 \Rightarrow$ we get the exponential distribution.

(ii) When $\theta_2 = 1$, we get the gamma distribution.

(iii) When $\theta_3 = 1$, we get the Weibull distribution.

The log-normal is also the limiting case as $\theta_3 \to \infty$.

1.3.7 REGRESSION MODELS :

So far we have considered several survival distributions for modelling the survival experience of a homogeneous population. However, there are explanatory variables upon which failure time may depend. Therefore, it is of interest to consider generalisations of these models to take into account of concomitant information on the individuals sampled.

Consider a failure time $T > 0$ and suppose a vector $\underset{\sim}{x}=(x_1,x_2,...,x_s)$ of explanatory variables or covariates has been observed. $\underset{\sim}{x}$ may include both quantitative and qualitative variables. The principal problem is that of modelling and determining the relationship between t and $\underset{\sim}{x}$.

1.3.7.1 Exponential Regression Models :

The exponential distribution can be generalised to obtain a regression model by allowing the failure rate to be a function of the covariates $\underset{\sim}{x}$. The hazards at time t for an individual with covariates $\underset{\sim}{x}$ is

$$\lambda(t;\underset{\sim}{x}) = \lambda(\underset{\sim}{x})$$

Thus the hazard for a given $\underset{\sim}{x}$ is a constant but the failure rate depends on $\underset{\sim}{x}$. Suppose the effect of the components of $\underset{\sim}{x}$ is linear. Then

$$\lambda(t;\underset{\sim}{x}) = \lambda c(\underset{\sim}{x}\ \underset{\sim}{\theta})$$

where $\underset{\sim}{\theta}' = (\theta_1,\theta_2,...\theta_s)$ is a vector of regression parameters, λ is a constant and c is a specified functional form. The choice of c may depend on the particular data being considered. Three specific forms have been used :

(1) $c(x) = 1 + x$

(2) $c(x) = (1+x)^{-1}$

(3) $c(x) = e^{x}$

The first two of these correspond to (1) the failure rate, (2) the mean survival time being linear function of $\underset{\sim}{x}$. They both suffer from the disadvantage that the set of $\underset{\sim}{\theta}$ values considered must be restricted to guarantee $c(\underset{\sim}{x}\ \underset{\sim}{\theta}) > 0$ for all possible $\underset{\sim}{x}$.

In many ways (3) is the most natural form since it takes only positive values. We use the form $c(x) = e^{x}$ here. Consider then the model with hazard function

$$\lambda(t;\underset{\sim}{x}) = \lambda e^{\underset{\sim}{x}\underset{\sim}{\theta}} \qquad \text{......(1.41)}$$

The conditional density function of T gives $\underset{\sim}{x}$ is then

$$f(t;\underset{\sim}{x}) = \lambda e^{\underset{\sim}{x}\underset{\sim}{\theta}}\ e^{-\lambda t e^{\underset{\sim}{x}\underset{\sim}{\theta}}}$$

In other words, the model (1.41) specifies that the log failure rate is linear function of covariates $\underset{\sim}{x}$. In terms of the log survival time Y = log T, the model (1.41) can be written as

$Y=\alpha-\underset{\sim}{x}\underset{\sim}{\theta}+W$ where $\alpha = -\log \lambda$ and W has extreme value distribution.

1.3.7.2 Weibull Regression Model :

In this model, the conditional hazard function is

$$\lambda(t;\underset{\sim}{x}) = \lambda\rho\ (\lambda t)^{\rho-1}\ e^{\underset{\sim}{x}\underset{\sim}{\theta}} \qquad \text{......(1.42)}$$

The conditional density of T is given as

$$f(t;\underset{\sim}{x}) = \lambda\rho\ (\lambda t)^{\rho-1}\ e^{\underset{\sim}{x}\underset{\sim}{\theta}}\ \ e^{-(\lambda t)^{\rho}e^{\underset{\sim}{x}\underset{\sim}{\theta}}} \text{......(1.43)}$$

The effect of the covariates is again to act multiplicatively on the Weibull hazard. If Y = log T, the model (1.43) is the linear model

$$Y = \alpha + \underset{\sim}{x}\underset{\sim}{\theta}^{*} + \sigma W \qquad \text{......(1.44)}$$

where $\alpha = -\log \lambda, \sigma = \rho^{-1}, \underset{\sim}{\theta}^{*} = -\sigma\underset{\sim}{\theta}$.

1.4 : SOFTWARE RELIABILITY :

As Software form an important part of many critical missions such as space shuttles and important systems such as nuclear reactors and heart monitors, the reliable operation of these projects/systems depends critically on the reliable operation of their software, the concept of Software Reliability has gained considerable importance.

The Life Cycle (LC) of software involves a series of production activities and can generally be divided into four phases Design, Coding, Testing and Operation/Maintenance. In spite of great advancement of the programming technology, the chances of error/fault occurrence due to human imperfection at every step are many. In other words software can never be made errors/fault/bug free. A fault/error is the defect in the software that, when executed under particular conditions, causes a failure, where failure means that the program in its functioning has not met users requirements is some way. To remove these faults/errors the software is tested under a large number of representative test cases with the intention of exposing faults/errors possibly contained in the program.

When a failure is observed, the code is in software to find the fault/error which caused the failure and an attempt is made to fix it. As a result reliability of the software is expected to increase. Thus it is very important to know the number of errors/faults remaining in the software or time interval between software failures. One of the approaches to software reliability is to describe a Software Error Detection Process which represents the behaviour of errors/faults during testing phase of the software development. Many models have been developed which attempt to estimate the errors content of a software and to predict the software reliability. These models are called Software Reliability Growth Model (SRGM). Since due to testing of a software, its reliability is expected to grow.

Here we first review some of the models based on Non-Homogeneous Poisson Process (NHPP), Models discussed here are either based on failure intensity function or the mean value function. The two are not necessarily interchangeable. In Chapter Two and Chapter Three of the thesis we developed two discrete software reliability growth models viz.

(i) A Discrete Software Reliability Growth Model with leading and dependent errors.

(ii) A Discrete Imperfect Debugging Software Reliability Growth Model.

Here we give a brief review of the release policies and different models.

1.4.1 ASSUMPTIONS :

Some of the general assumptions assumed in every model are as follows :

1. Software System is subject to failure during execution caused by errors/faults remaining in the system.
2. Failure rate of the software is equally affected by errors/faults remaining in the software.
3. The number of failures detected at any time is proportional to the remaining number of errors/faults in the software.
4. On a failure detected at any time is proportional.
5. All faults/errors are mutually independent from failure detection point of view.
6. The proportionality of failure detection/fault isolation/fault removal is constant.
7. Corresponding to the error detection/removal phenomenon at the manufacturer/user end, there exists an equivalent error detection/removal phenomenon at the user/manufacturer end.

8. Software Life Cycle is more than the optimum release time.

9. The error detection/removal phenomenon is modelled by NHPP.

1.4.2 NOTATIONS :

$[N(t);\ t>0]$ ≡ Counting process representing the cumulative number of failures/isolation/removals in $(0,t)$.

a ≡ S-expected initial errors/fault context.

b ≡ Error detection/isolation/removal rate per error.

b_i ≡ Initial value of b.

b_f ≡ Final value of b.

$m_f(t)$ ≡ S-expected number of failures in $(0,t)$.

$m_i(t)$ ≡ S-expected number of isolations in $(0,t)$.

$m_r(t)$ ≡ S-expected number of removals in $(0,t)$.

$b(t)$ ≡ Error-detection rate per error as a function of t.

$b(m_f)$ ≡ Failure detection rate per error as a function of $m_f(t)$.

$\lambda(t)$ ≡ Failure intensity at t.

$w(t)$ ≡ Current testing effort expenditure at time t

$W(t)$ ≡ Cumulative testing effort by time, t.

$$\equiv \int_0^t w(x)\, dx.$$

α, β ≡ are parameters of testing effort function.

$R(x|t)$ ≡ Reliability of a Software in $(t, t+x)$.

λ_o ≡ Initial failure intensity.

T_{IC} ≡ Life Cycle of the Software.

C_1 ≡ Cost of removing a fault before releasing.

C_2 ≡ Cost of removing an error after releasing.

C_3 ≡ Testing Cost per Time.

$C(T)$ ≡ Total Cost Function.

λ_d ≡ Desired level of failure intensity to be achieved.

R_d ≡ Desired level of reliability to be achieved.

SOME IMPORTANT DEFINITIONS

1.4.3 NON-HOMOGENEOUS POISSON PROCESS (NHPP) :

Let $\{N(t);\ t \geq 0\}$ be a counting process representing the cumulative number of failures by time t. $N(t)$ is a random variable and the process $\{N(t);\ t \geq 0\}$ is NHPP if

(i) $N(0) = 0$

(ii) $\{N(t);\ t \geq 0\}$ has independent increments.

(iii) P_r[two or more events in $(t, t+h)$] $= O(h)$.

(iv) P_r[exactly one event in $(t, t+h)$] $= \lambda(t)h + O(h)$. Where $\lambda(t)$ is the intensity function of $N(t)$.

and if we let $m(t) = \int_0^t \lambda(x)\, dx$ represent $m_f(t)$ or $m_i(t)$ or

$m_r(t)$ depending upon the model, then it can be shown that

$$P_r[N(t)=n] = \frac{[m(t)]^n e^{-m(t)}}{n!}; \; n=0,1,2,\ldots \qquad \ldots\ldots(1.45)$$

i.e. N(t) has poisson distribution with expected value $E[N(t)]=m(t)$ for $t > 0$ and m(t) is called the mean value function of the NHPP.

1.4.4 SOFTWARE RELIABILITY GROWTH MODEL :

Software Reliability Growth Model is defined as a mathematical relation between the time span of testing (or using) the Software and the cumulative number of errors/faults detected.

1.4.5 SOFTWARE RELIABILITY :

Software Reliability is defined as the probability that a software failure does not occur in $[t,t+x]$ given that the most recent failure occurred at time $t>0$, $x>0$. It can be shown that for a model based on NHPP

$$R(x|t) = e^{-[m_f(t+x)-m_f(t)]} \qquad \ldots\ldots(1.46)$$

1.4.6 FAILURE INTENSITY :

The Failure Intensity function is the state of change of the mean value function representing the average cumulative number of failures associated with the given time point.

i.e. $$\lambda(t) = \frac{d}{dt}[m_f(t)] \qquad \ldots\ldots(1.47)$$

It may also be defined as the number of failures per unit time.

1.4.7 MODELS BASED ON FAILURE INTENSITY :

Starting with the failure intensity function. Musa proposed two models (Second with Okumoto) through which it is attempted to predict future failure behaviour and reliability of software.

1.4.7.1 Basic Execution Time Model [Musa] :

It has a failure intensity function which decays exponentially with execution time t

i.e. $\lambda(t) = \lambda_o e^{-bt}$ ......(1.48)

where b is the rate of decrease per time (which is same as error detection rate per error). The initial failure intensity λ_o is given as

$\lambda_o = ab$ ......(1.49)

The expression for the expected number of failures by time t can be obtained from (1.48) as

$m_f(t) = a(a-e^{-bt})$ ......(1.50)

and the failure intensity can be expressed as a function of $m_f(t)$ as

$\lambda(m_f) = b\ [a-m_f(t)]$ ......(1.51)

(1.51) shows that the rate of change of failure intensity with respect to failures experienced is constant whether it is the first failure at the last that is being fixed.

1.4.7.2 Logarithmic Poisson Model (Musa and Okumoto) :

This model has a failure intensity function which decays exponentially with respect to the mean value function $m_f(t)$

i.e. $\lambda(t) = \lambda_o \; e^{-bm_f(t)}$ ......(1.52)

where b is rate of decrease per failure. From equation (1.52), we get

$m_f(t) = \log_e \; (1+\lambda_o bt)/b$ ......(1.53)

The "Logarithmic Poisson Model" is derived from the form (1.53). The failure intensity as a function of $m_f(t)$ is given as

$\lambda(m_f) = \lambda_o[e^{-b}] \; m_f(t)$ ......(1.54)

This show that the rate of change of intensity with respect to failures experienced decreases exponentially with failures experienced and is not constant as in the case of basic execution time model. It means that the first failure initiates a repair process that yields a substantial decrease in failure intensity, while later failure results in much smaller decrements. This is because during testing, the software is first tested on frequently used inputs.

Moreover it may be noted that by time infinity, the failure intensity reduces to zero and the number of failures experienced is infinity. This is possible when either at the time of debugging faults are being introduced or the debugging process is imperfect or each fault generates more than one failure or when combination of these are possible. It is obvious in this case that the parameters of interest are not the number of errors in a software, but the failure intensity and the rate at which failures are occurring.

1.4.8 EXPONENTIAL MODELS BASED ON MEAN VALUE FUNCTION :

This category of SRGMs is based on slightly different assumption models are formulated based on an expression for the mean value function of the Poisson Process rather than the failure intensity function and are suitable for finite error content.

1.4.8.1 Goel-Okomoto Model (Goel & Okomoto (1979)) :

Assuming that the S-expected number of failures in $(t,t+\Delta t)$ is essentially proportional to the S-expected number of undetected errors at time t, the following equations can be easily written

$$m_f(t+\Delta t) - m_f(t) = b(a-m_f(t))\ \Delta t+O(\Delta t) \qquad ......(1.55)$$

where $O(\Delta t)/\Delta t \longrightarrow 0$ as $\Delta t \longrightarrow 0$. By letting $\Delta t \longrightarrow 0$ in (1.55) gives

$$m_f(t) = b[a-m_f(t)] \qquad ......(1.56)$$

(1.56) together with $m_f(0) = 0$ gives

$$m_f(t) = a(1-e^{-bt}) \qquad ......(1.57)$$

which is mathematically isomorphic to the mean value function of Basic Execution Time Model. Here it is assumed that the fault causing failure is immediately removed and hence fault detection and failure occurrence are assumed to be synonymous.

1.4.8.2 Exponential Model with Imperfect Debugging (Kapur and Garg (1990)) :

During testing phase of a software on failure an attempt is made to correct the cause of the failure. However, it is not always possible to find the cause of the failure and remove it. This may be attributed to lack of sufficient knowledge about the software, poor documentation of the software and so on.

In this model it is assumed that on failure instantaneous repair effort starts and the following may occur :

(i) fault content is reduced by one with probability p_o.

(ii) fault content is unchanged, with probability $1-p_o$.

Based on these assumptions, following different equations may be written as

$$m_f'(t) = b\ p_o\ [1-m_r(t)] \qquad \ldots\ldots(1.58)$$

This gives

$$m_r(t) = a\ [1-e^{-bp_o t}] \qquad \ldots\ldots(1.59)$$

and

$$m_f(t) = \frac{a}{p_o}\ [\ 1-e^{-bp_o t}] \qquad \ldots\ldots(1.60)$$

(1.60) implies that the expected number of failures by infinite time will be greater that 'a' which is because that some faults may not be removed even though an attempt was made to do it.

1.4.8.3 Exponential Model with Testing Effort (Yamada et al (1986)) :

In general some resources like, manpower, C.P.U. time etc. are spent during testing phase of the software development. The consumption curve of testing resources over the testing period can be thought of as a testing effort curve.

In this model it is assumed that the S-expected number of errors detected in the time interval $(t,t+\Delta t)$ to the current testing effort expenditure is proportional to the S-expected number of remaining errors. So,

$$\frac{m_f'(t)}{w(t)} = b[a-m_f(t)] \qquad \ldots\ldots(1.61)$$

This gives

$$m_f(t) = a[1-e^{-bw(t)}] \qquad \ldots\ldots(1.62)$$

It is generally assumed that testing effort in the software development process follows exponential or Reyleigh Curve i.e.

$$W(t) = \alpha\ [1-e^{-\beta t}] \qquad \ldots\ldots(1.63)$$

or

$$W(t) = \alpha\ [1-e^{-\beta t^2/2}] \qquad \ldots\ldots(1.64)$$

Values of α and β are estimated respectively from

(1.63). These estimated values of α and β are then put in (1.64) to get the estimates of a and b.

Since resources are finite, $ae^{-\alpha b}$ number of faults will remain undetected by time infinity.

1.4.9 S-SHAPED SOFTWARE RELIABILITY GROWTH PHENOMENON :

It is generally accepted that there exists S-shaped software reliability growth phenomenon that is observed in the testing phase of the development of the softwares. It has been interpreted in different ways. Ohba (1984) interprets it as initial delay of fault isolation after the initial failure detection. Yamada (1984) interprets it as unskilledness of the testing team. Ohba (1984) developed a SRGM for a situation where a fault in a software is dependent on the previous fault detected and the cumulative number of failures/faults detected curve of this model also shown S-shaped phenomenon. This same S-shaped phenomenon is again observed when the testing effort curve of Yamada et al (1986) follow Rayleigh type of distribution. Again in flexible model of Bittanti (1988) this phenomenon is observed when the error detection rate per error is increasing.

1.4.9.1 Delayed S-shaped SRGM (Ohba (1984)) :

Fault removal, in this model is assumed to be two phase process consisting of failure detection and its eventually removal by isolation. It takes care of the time taken to isolate and remove a fault and so it is important that the data to be used here should be that of fault isolation.

It is further assumed that the number of faults isolated any time is proportional to the current number of faults not isolated. Failure rate and isolation rate per error are assumed to be same and equal to b. Thus

$$m_f'(t) = b[a-m_f(t)] \quad ....(1.65)$$

$$m_r'(t) = b[m_f(t) - m_r(t)] \quad ....(1.66)$$

Solving these we get, mean value function as

$$m_r(t) = a[1-(1+bt)e^{-bt}] \quad ....(1.67)$$

Thus this model is called S-shaped because the graph of cumulative number of fault removed with respect to time has a jump at the initial portion of the graph this model can be further be extended depending upon the severity of the error in the software.

1.4.9.2 Inflection S-shaped SRGM(Ohba (1984)) :

The mean value function of this model is given by

$$m_f(t) = \frac{1-e^{-bt}}{1-\phi e^{-bt}} \quad ....(1.68)$$

where ϕ is called inflection parameter is equal to $\frac{1-r}{r}$ where r is the ratio of the number of detectable faults to the total number of faults.

This model take care of situation where error/faults are mutually dependent and is based on the assumption that the error detection rate per error increases throughout the test period.

1.4.9.3 <u>Exponential with Rayleigh Type Testing Effort (Yamada et al (1986))</u> :

As discussed in (1.4.8.), its mean value function is given by

$$m_f(t) = a\,[1-e^{-bw(t)}] \qquad \ldots\ldots(1.69)$$

And when the testing effort curve follows a Rayleigh curve given by

$$W(t) = \alpha[1-e^{-\beta t^2/2}] \qquad \ldots\ldots(1.70)$$

Then the reliability growth phenomenon is S-shaped.

1.4.10 DISCRETE SOFTWARE RELIABILITY GROWTH MODELS :

In this class of SRGMs the number of test runs or the number of executed test cases is taken as the unit of error detection period. Random variable of the NHPP is defined as the number of errors detected/removed by n test runs.

For this class of SRGMs assumptions are modified in the discrete sense. Little work has been done in this class of SRGMs.

1.4.10.1 <u>Error Content Proportional Detection Rate Model (Yamada et al (1985))</u> :

It is assumed that the expected number of failures between nth and (n + 1)th executed test cases is proportional to the expected number of faults remaining in it i.e.

$$m_f(n+1) - m_f(n) = b\,[a-m_f(n)] \qquad \text{......(1.71)}$$

Solving the difference equation, we get

$$m_f(n) = a[1-(1-b)^n] \qquad \text{......(1.72)}$$

and $$\lambda(n) = ab(1-b)^n \qquad \text{.....(1.73)}$$

1.4.10.2 Geometric Error Detection Rate SRGM(Yamada et al(1985)):

This SRGM is developed under the assumption that the ratio of the error detection rate for any test run and the rate for its predecessor is constant less than unity, and the expected number of failures per test is geometrically decreasing, So

$$m_f(n+1) - m_f(n) = Dr^n \qquad \text{......(1.74)}$$

where r is decreasing ratio for the expected number of error detected per test run and D is the initial expected number of errors detected by the first test run. Thus

$$m_f(n) = D\frac{1-r^n}{1-r} \qquad \text{......(1.75)}$$

And the expected number of errors to be eventually detected i.e the expected initial error content is given by

$$m_f(\infty) = D/(1-r) \qquad \text{......(1.76)}$$

1.4.10.3 Discrete S-shaped SRGM (Kapur et al (1990)) :

This model takes care of the delay between the failure caused by a fault and its subsequent detection. It is a two phase phenomenon and so the mean value function of the model is given by

$$m_r(n) = a[1-(1+bn)(1-b)^n] \qquad \text{......(1.77)}$$

1.4.11 ESTIMATION OF PARAMETERS OF SRGM :

Our main aim in software reliability modelling is to estimate different model parameters so that the future behaviour of the software can be predicted suppose that the data on n failure times (or isolation or removal times) $S(s_1,s_2,...,s_n)$ where $s_1 < s_2 < ... < s_n$ are observed during testing. Then the likelihood function for unknown parameters of the model given S is given by

$$L = e^{-m(S_n)} \prod_i \in (S_i) \qquad \text{......(1.78)}$$

where

$$\in(S_i) = \frac{d}{dt} (m(t)) \text{ at } t = S_i \qquad \text{......(1.79)}$$

where m(t) is the mean value function of the underlaying NHPP. From (1.78) maximum likelihood estimates of parameters can easily be obtained.

Alternatively, suppose that the data on cumulative number of failures (or isolation or removals $Y_k(O)$ $(0 < y_1 < y_2 <...< y_n)$ in a given time interval $[o,t_k]$ are observed. Then the likelihood function of the unknown parameters is given by

$$L = \prod_{k} \frac{[m(t_k) - m(t_{k-1})]^{Y_k - Y_{k-1}}}{[Y_k - Y_{k-1}]} e^{-[m(t_k)-m(t_{k-1})]} \qquad ..(1.80)$$

From (1.80), estimates of unknown parameters can easily be obtained.

1.4.12 PREDICTIVE VALIDITY OF MODELS (MUSA ET AL (1989)) :

It is ability of the model to determine future failure/removal behaviour during either the test or the operational phase from present and past failure/removal behaviour in the respective phase. It is very effective tool to compare the applicability of models on a particular data.

Here we attempt to predict the number of failures/removals that will be experienced by the end of the period of testing over which the data has been collected and compared this with actual valves. Assume that n failures/removals has been observed by the end of the times S_n where S_i is the time to ith failure/removal.

Clearly, $0 < S_1 < S_2 < .... < S_n$ the failure/removal data upto time $S_t (<S_n)$ is used to estimate the parameters of the mean valve function $m_f(t)$ or $m_r(t)$. Then the number of failures removals by time S_n can be predicted by substituting the estimates of parameters in $m_f(t)$ which is compared with the actually observed number

n. This is repeated for various values of S_t.

The predictive validity can be checked visually by ploting normalised relative number of errors or removals $[m_f(S_n)-n]/n$ or $[m_r(S_n)-n]/n$ against the normalised time S_t/S_n. The error/removal will approach to zero as at approaches S_n. If the points are positive (negative) the model tends to overestimate (underestimate). Number closer to zero implies more accurate prediction and hence the better model.

1.4.13 OPTIMAL RELEASE POLICY :

It is of utmost importance to find the appropriate release time of the software. If the release of the software is unduly delayed, manufacturer may, suffer in terms of penalties and revenue loss, while a premature release may cost heavily in terms of the fixes to be done after release and may even harm manufactures reputation. Therefore, manufacturer must have some idea about the possible attributes of the softwares like its initial error contents failure rate, reliability at time t and its potential release time. Thus it is of interest to know when to stop testing and realize the software. Software release time problem has been classified in different ways.

One is, when t release a software so that the cost incurred during the life cycle of the software ie during development and operational phases is minimised or reliability reaches a desired level.

Software cost includes the cost of removing errors both before and after release of the software and testing cost during testing phase. Assuming that the software is released at time T cost function in general can be written as

$$C(T) = C_1 m_r(T) + C_2(m_r(T_k) - m_r(T)) + C_3 T \quad \ldots.(1.81)$$

Thus the problem of finding optimal release time reduced

$$\min C(T)$$

subject to

$$R(x|T) > R_d \text{ or } \lambda(T) < \lambda d \quad \ldots\ldots(1.82)$$

$$T > 0$$

For the discrete models, T is replaced by n, the number of executed test cases.

Alternatively, this problem can be redefined in terms of maximizing gain, which is defined as the difference in cost incurred when all the errors are removed during operational phase as against the cost, when some errors are removed during the testing phase and others are removed during the operational phase. thus denoting it by G(T), we have

$$G(T) = (C_2 - C_1)\ m_r(T) - C_3 T \quad \ldots\ldots(1.83)$$

It can be seen that maximizing gain is same as minimizing cost.

We have so far assumed that the software life cycle is constant. However, in real situation it might be more appropriate to assume it to be random variable because software system may be abandon ed as new versions are available. Recently Yun and Bai [ ] discussed a software release policy maximizing profit alone, when the price of the software and the expenditure incurred on testing, correcting and error during testing and operation. One may also incorporate the idea of penalty cost which in incurred by the manufacturer by not delivering the software by scheduled delivering time.

1.5 : THE PROPORTIONAL HAZARDS MODEL :

Let $\lambda(t;x)$ represent the hazard function at time t for an individual with covariates x.

The proportional hazards model due to Cox(1972) specifies that

$$\lambda(t;\underset{\sim}{x}) = \lambda_o(t)\, e^{\underset{\sim}{x}\underset{\sim}{\beta}} \qquad \ldots\ldots(1.5.1)$$

where $\lambda_o(t)$ is an arbitrary unspecified base line hazard function for continuous T and $\underset{\sim}{x}$ is a row vector a k measured variates $\underset{\sim}{\beta}$ is a column vector of k regression parameters, T is the associated failure time.

In this model, the covariates act multiplicatively on the hazard function. If $\lambda_o(t) = \lambda$, the model (1.5.1) reduces to the exponential regression model

$$\lambda(t;\underset{\sim}{x}) = \lambda e^{\underset{\sim}{x}\underset{\sim}{\beta}} \qquad \text{......(1.5.2)}$$

If $\lambda_o(t) = \lambda p(\lambda t)^{p-1}$ then the model reduces to

$$\lambda(t;\underset{\sim}{x}) = \lambda p(\lambda t)^{p-1}\, e^{x\underset{\sim}{\beta}} \qquad \text{......(1.5.3)}$$

Corresponding to model (1), the conditional density function of T given $\underset{\sim}{x}$ is

$$f(t;\underset{\sim}{x}) = \lambda_o(t)\; e^{x\underset{\sim}{\beta}}\; e^{-\left\{e^{-\underset{\sim}{x}\underset{\sim}{\beta}} \int_o^t \lambda_o(u)\, du\right)}$$

$$= \lambda(t;\underset{\sim}{x})\; S(t;\underset{\sim}{x}) \qquad \text{......(1.5.4)}$$

The conditional survivor function for T given $\underset{\sim}{x}$ is

$$S(t;\underset{\sim}{x}) = \Big[S_o(t)\Big]^{\underset{\sim}{x}\underset{\sim}{\beta}} = e^{-\int_o^t \lambda_o(u)\; e^{\underset{\sim}{x}\underset{\sim}{\beta}}\; dx} \qquad \text{......(1.5.5)}$$

where $\qquad S_o(t) = e^{-\int_o^t \lambda_o(u)du}$

Thus the survivor function of t for a covariate value $\underset{\sim}{x}$ is obtained by raising the base line survivor function $F_o(t)$ to a power.

The two important generalisation that do not substantially complicate the estimation of $\underset{\sim}{\beta}$. First, the nuisance function $\lambda_o(t)$ can be allowed to vary in specific subsets of the data. Suppose the data is divided into k strata and that hazard $\lambda_j\ (t;\underset{\sim}{x})$ in the jth stratum depends on an arbitrary function $\lambda_{oj}(t)$ and can be written

$$\lambda_j(t;\underset{\sim}{x}) = \lambda_{oj}(t)\; e^{\underset{\sim}{x}\underset{\sim}{\beta}} \qquad \text{......(1.5.6)}$$

for $j = 1,2,\ldots,k$

Such generalisation is useful when some explanatory variable or variables do not appear to be multiplicative effect on the hazard function. The range of such variables can then be divided into start with only the remaining regression variables contributed to the exponential factor in (1.5.6). The second important generalisation allows the regression variables $\underset{\sim}{x}$ to depend on time itself. Such regression variables arise in the heart transplant. Where treatment group itself is time dependent as are certain donor recipient matching variables.

1.5.1 THE ACCELERATED FAILURE TIME MODEL :

The model specified in (1.5.1) is the is the multiplicative effect of regression variables on the hazard function. This model does not postulate direct relationship between $\underset{\sim}{x}$ and t. As we know that the exponential and Weiball regression models are linear in

$$Y = \log T$$

where the conditional density function of given $\underset{\sim}{x}$ is

$$-f(t;\underset{\sim}{x}) = \lambda e^{\underset{\sim}{x}\underset{\sim}{\theta}} e^{-\lambda t e^{\underset{\sim}{x}\underset{\sim}{\theta}}}$$

$$\log f(t;x) = \log \lambda + x\underset{\sim}{\theta} - \lambda t e^{\underset{\sim}{x}\underset{\sim}{\theta}}$$

$$= \alpha - \underset{\sim}{x}\theta + W$$

Here we obtain a second class of survival models, the class of log linear models for T.

Suppose that Y = log T is related to the covariates $\underset{\sim}{x}$ via a linear model

$$\underset{\sim}{Y} = x\underset{\sim}{\theta} + U \qquad \text{......(1.5.7)}$$

where U is an error variable with density f.

Exponentiation gives

$$T = e^{\underset{\sim}{Y}} = e^{\underset{\sim}{x}\theta+u} = e^{\underset{\sim}{x}\underset{\sim}{\theta}} e^{u}$$

$$= T'e^{\underset{\sim}{x}\underset{\sim}{\theta}} \qquad \text{......(1.5.8)}$$

where $T'=e^{w} > 0$ has hazard function $\lambda_o(t)$ that is independent of $\underset{\sim}{\theta}$,

The hazard function for T can be written in terms of base line hazard $\lambda_o(t)$.

$$\lambda(\underset{\sim}{t};\underset{\sim}{x}) = \lambda_o(\underset{\sim}{t}e^{-\underset{\sim}{x}\underset{\sim}{\theta}})\, e^{-\underset{\sim}{x}\underset{\sim}{\theta}} \qquad \text{......(1.5.9)}$$

The survivor function is

$$S(\underset{\sim}{t};\underset{\sim}{x}) = e^{-\int_o^{te^{-x\theta}} \lambda_o(u)\, du} \qquad \text{......(1.5.10)}$$

and the probability density function is the product of equation (1.5.9) and equation (1.5.10) as

$$f(t;\underset{\sim}{x}) = \lambda(t;\underset{\sim}{x})\; S(t;\underset{\sim}{x})$$

$$= \lambda_o\left[t\bar{e}^{\underset{\sim}{x}\underset{\sim}{\theta}}\right] e^{-x\underset{\sim}{\theta}}\, e^{-\int_o^{te^{-x\theta}} \lambda_o(u)\, du} \qquad \text{...(1.5.11)}$$

This model specifies that the effect of the covariable is multiplicative of t rather than on the hazard function. That is, we assume a base line hazard function to exist and that the effect of the regression variables is to alter the rate at which an individual proceeds along the time axis. It is supposed that the role of $\underset{\sim}{x}$ is to accelerate the time to failure. The model (1.5.9) is known as Accelerated Failure

Time Model.

All the parametric models lead to linear models for $\underset{\sim}{Y}$. The exponential and Weibull regression models can be considered as special cases of either the Accelerated Failure Time Model or the Proportional Hazards Model. However, the log linear models derived from other parametric models are not special case of the proportional hazard models. For example, log-normal hazard function with different location parameters α_1 and α_2 are generally not proportional to one another.

1.5.2 COMPARISON OF PROPORTIONAL HAZARD MODELS AND ACCELERATED FAILURE TIME MODELS :

Let us consider the intersection of the proportional hazard models and the Accelerated Failure Time Models. Consider the subset of log-linear models in which the regression variable acts multiplicative on hazard function. Consider the proportional hazard model

$$\lambda(t;\underset{\sim}{x}) = \lambda_{01}(t)\, e^{\underset{\sim}{x}\underset{\sim}{\theta}} \qquad \text{for all } \underset{\sim}{t},\underset{\sim}{x}$$

and the accelerated failure time model

$$\lambda(t;\underset{\sim}{x}) = \lambda_{02}\left[te^{\underset{\sim}{x}\theta_2}\right] e^{\underset{\sim}{x}\theta_2} \qquad \text{for all } \underset{\sim}{t}\ \&\ \underset{\sim}{x}$$

$$\text{i.e.} \quad \lambda_{01}(t)\, e^{\underset{\sim}{x}\underset{\sim}{\theta}} = \lambda_{02}\left[te^{\underset{\sim}{x}\theta_2}\right] e^{\underset{\sim}{x}\theta_2}$$

The value $\underset{\sim}{x} = 0$ gives $\lambda_{01}(.) = \lambda_{02}(.) = \lambda_{0}(.)$

while $\underset{\sim}{x} = (-\log t/\theta_{21},0,....,0)$ gives at that t

$$\lambda_o(t)t^{\theta_{11}\theta_{12}^{-1}} = \lambda_o(1)t^{-1}$$

where θ_{11} and θ_{21} are the first component of $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$

$$\underset{\sim}{\theta}_1 = \begin{bmatrix} \theta_{11} \\ \theta_{12} \\ \vdots \\ \theta_{1k} \end{bmatrix} \qquad \underset{\sim}{\theta}_2 = \begin{bmatrix} \theta_{21} \\ \theta_{22} \\ \vdots \\ \theta_{2k} \end{bmatrix}$$

It follows that for all t

$$\lambda_o(t) = \lambda p(\lambda t)^{p-1}$$

where $p = \theta_{11}\theta_{21}^{-1}$

$\lambda = \left\{\lambda_o(1)/p\right\}^{1/p}$ Note that $\underset{\sim}{\theta}_1 = -p\underset{\sim}{\theta}_2$

The Weibull and Exponential log-linear models are then the only log-linear models m (1.5.1). This leads to a characterization of the parameter Weibull model as the unique family that is closed under both multiplication on failure time and multiplication of the hazard function by an arbitrary non-zero constant.

1.6 : DISCRETE FAILURE TIME MODELS :

The models so far discussed are appropriate for failure time data arising from continuous distributions. However, failure time data is discrete which arises either through the grouping of continuous data due to imprecise measurement or because the time itself is discrete.

Let x has a Weibull distribution with survivor function

$$S(x) = e^{-(\lambda x)^p} \qquad \text{......(1.6.1)}$$

and times are grouped into unit intervals so that the discrete observed variable is

T = [X] where [X] represents "integer part of X".

The probability function of T is given by

$$p(t) = P(T=t) = P[t \leq X \leq t+1]$$
$$= P(X \geq t) - P(X > t+1)$$
$$= \theta^{tp} - \theta^{(t+1)p}; \quad t = 1,2,.. \qquad ......(1.6.2)$$

where $\theta = e^{-\lambda^p}$, $0 < \theta < 1$.

The special case p=1 is the geometric distribution with probability function $\theta^t(1-\theta)$.

i.e. $p(t) = \theta^t \ (1-\theta) \ ; \ t = 0,1,2$

The hazard function corresponding to (1.6.2)

$$\lambda(t) = P[T = t \mid T \geq t]$$
$$= 1 - \theta^{(t+1)p-tp} \qquad ......(1.6.3)$$

(i) $\lambda(t)$ is monotonic increasing for $p > 1$.

(ii) $\lambda(t)$ is monotonic decreasing for $p < 1$.

(iii) $\lambda(t)$ constant for p=1.

This can be generalised to a regression model by applying the same grouping to the Weibull Regression Model.

1.6.1 DISCRETE PROPORTIONAL HAZARDS MODEL :

Let the failure time T given covariates $\underset{\sim}{x}$ have a discrete distribution with mass point at $0 \leq x_1 < x_2 < ...$ and so on. Let $S_0(t)$ represent the base line survivor function for $\underset{\sim}{x} = \underset{\sim}{0}$. The corresponding survivor function for covariates $\underset{\sim}{x}$ is

$$S(t,\underset{\sim}{x}) = [S_o(t)]^{e^{\underset{\sim}{x}\underset{\sim}{\theta}}} \qquad \ldots\ldots(1.6.4)$$

If the hazard function corresponding to S_o has contribution λ_i at x_i then

$$S_o(t) = \prod_{i \mid x_i < t} (1 - \lambda_i)$$

and

$$S(t,\underset{\sim}{x}) = \prod_{i \mid x_i < t} (1 - \lambda_i)^{e^{\underset{\sim}{x}\underset{\sim}{\theta}}} \qquad \ldots\ldots(1.6.5)$$

The hazard at x_i for covariate $\underset{\sim}{x}$ is then

$$1 - (1-\lambda_i)^{e^{\underset{\sim}{x}\underset{\sim}{\theta}}} \qquad \ldots\ldots(1.6.6)$$

Discrete model (1.6.6) can also be obtained by grouping the continuous model. If the continuous failure time arising from the proportional hazard model are grouped into disjoint intervals

$$[0 = \alpha_o,\alpha_1),\ [\alpha_1,\alpha_2),\ldots,[\alpha_{k-1},\alpha_k = \infty)$$

The hazard of failure in the interval for an individual with covariate $\underset{\sim}{x}$ is

$$P\left[T\in[\alpha_{i-1},\alpha_i) \mid T \geq \alpha_{i-1}\right] = \left(1-(1-\lambda_i)^{e^{\underset{\sim}{x}\underset{\sim}{\theta}}}\right) \ldots\ldots(1.6.7)$$

where $$\lambda_i = e^{-\int_{\alpha_{i-1}}^{\alpha_i} \lambda_o(u)\, du}$$

This discrete model is then the uniquely appropriate one for grouped data from the continuous proportional hazard model.

If the discrete base line hazard function is given by

$$\lambda_d(t)\ dt = \sum \lambda_i\ \delta(t-x_i)\ dt \qquad \text{......(1.6.8)}$$

We see that the hazard function for covariates $\mathbf{x}$ is

$$\lambda(t;\mathbf{x})\ dt = 1 - (1 - \lambda_d(t)\ dt)^{e^{\mathbf{x}\boldsymbol{\beta}}} \qquad \text{......(1.6.9)}$$

It $\lambda_d(t)$ is replaced with a continuous hazard $\lambda(t)$ in (1.6.9) the relationship is precisely

$$\lambda(t;\mathbf{x}) = \lambda_o(t)e^{\mathbf{x}\boldsymbol{\beta}}$$

Thus if $\lambda_o(t)$ is the base line hazard function $\mathbf{x} = 0$ for a discrete or continuous or mixed random variables. The relationship between the survivor and hazard function is

$$S(t;\mathbf{x}) = \mathop{\mathcal{P}}_{o}^{t}\ [1 - \lambda(u;\mathbf{x})\ du]$$

$$= \mathop{\mathcal{P}}_{o}^{t}\ [1 - \lambda_o(u)\ du]^{e^{\mathbf{x}\boldsymbol{\beta}}}$$

where $\mathcal{P}$ is the product integral is defined as

$$\mathop{\mathcal{P}}_{o}^{t}\ [1 - d\Lambda(u)] = \lim \prod_{1}^{r} [1 - \{\Lambda(u_k)-\Lambda(u_{k-1})\}]$$

Here $0<u_o<u_1<\ldots<u_r=t$ and the limit is taken as $r \longrightarrow \infty$ and $u_k - u_{k-1} \longrightarrow 0$.

$$S(t) = \mathop{\mathcal{P}}_{o}^{t}\ [1 - \lambda(u)\ du]$$

1.7 : METHODS OF ESTIMATION OF PROPORTIONAL HAZARD MODELS :

Here our attention is focused on different methods of estimation of data arising from the proportional hazards model.

In parametric case the failure time distribution is assumed known except for a few scalar parameters. The proportional hazards model, however, is non parametric in the sense that it involves an unspecified function in the form of an arbitrary base-line hazard function. In consequence, this model is more flexible but different approaches are required for estimation.

The main problems addressed are those of estimation of β and $\lambda_o(.)$. The different methods of estimating the parameters β are

1. Method of Marginal Likelihood.
2. Method of Partial Likelihood.
3. Breslow's Maximum Likelihood Method.

1.7.1 METHOD OF MARGINAL LIKELIHOOD :

Suppose that n individuals are observed to fail at $t_1, t_2, ..., t_n$ with corresponding covariates $x_1, x_2, ..., x_n$. We assume that all failures are distinct. Let $O(t) = (t_1, t_2, ..., t_n)$ be the order statistics and $P(t)=((1),(2),...,(n))$ be the rank statistics, the order statistics refer to the $t_{(i)}$'s ordered from smallest to largest i.e. $t_{(1)} < t_{(2)} <....< t_{(n)}$ and the notation (i), in the rank statistics refers to the level attached.

Consider the model (1.5.1) and define $u = g^{-1}(t)$ where $g \in G$ the group of strictly increasing and differentiable transformation of $(0,\infty)$ onto $(0,\infty)$ the conditional distribution of u given x,has the hazard

$$\lambda(u,\theta|x) = \lambda_1(u)e^{x\theta}$$

where
$$\lambda_1(u) = \lambda_o g(u)\ g^1(u)$$

Thus, if the data were the presented in the form of $u_1,u_2,...,u_n$ and $x_1,x_2,...,x_n$ where $g(u_i) = t_i$ in inference problem about θ be the same provided $\lambda_o(.)$, were completely unknown. In effect the estimation problem for θ is invariant under the group G of transformation on the survival time t. For inference about θ, the marginal distribution of the rank's is available and the marginal likelihood is proportional to probability that the rank vector should be observed. That is the marginal likely is proportional to

$$P(\rho;\underset{\sim}{\theta}) = P(\rho = [(1),(2),...,(n)];\underset{\sim}{\theta})$$

$$= \int_0^\infty \int_{t_{(1)}}^\infty \cdots\cdots \int_{t_{(n-1)}}^\infty \prod_{i=1}^n f(t_{(i)};x_{(i)})\ dt_{(n)}...dt_{(1)}$$

$$= \frac{e^{\sum_{i=1}^n x_{\underset{\sim}{i}}\theta}}{\prod_{i=1}^n \left[\sum_{f\in R(t_{(i)})} e^{x_{\underset{\sim}{i}}\theta} \right]}$$

Where $R(t_{(i)})$ is the set of levels attached to the individuals at risk gust prior to $t_{(i)}$.

i.e. $R(t_{(i)}) = ((i),\ (i+1),\ \longleftarrow\ (n))$ some modification is required for handling censured data. If all the items are simultaneously put on test and follow to the K_{th} failure time (type II censoring), a marginal likelihood is again easily obtained.

In this case group acts transitively on the censoring time and the invariants in the sample space is the first k rank variables [(1),(2),...,(k)].

The argument could be extended to progressive type II censoring patterns where items are withdrawn from test with each failure.

1.7.2 METHOD OF PARTIAL LIKELIHOOD :

The method of Partial Likelihood proposed by Cox (1975) for the proportional hazard models.

Consider the set $R(t_{(i)},0)$ individual at risk at $t_{(i)}-0$, the conditional probability that item (i) fails at $t_{(i)}$ given that the items $R(t_{(i)})$ the exactly one failure occurs at $t_{(i)}$ is

$$\frac{\lambda(t_{(i)};x_{(i)})}{\sum\limits_{l\in R(t_{(i)})} \lambda(t_{(i)};x_l)} = \frac{e^{x_{(i)}\theta}}{\sum\limits_{l\in R(t_{(i)})} e^{x_l\theta}} \qquad ......(1.7.1)$$

for $i = 1,2,...,k$

The partial likelihood for θ is now formed by taking the product over all failure points $t_{(i)}$ of (1.7.1)

$$L(\theta) = \prod_{i=1}^{k} \left[\frac{e^{x_{(i)}\theta}}{\sum\limits_{l\in R(t_{(i)})} e^{x_l\theta}}\right] \qquad ....(1.7.2)$$

which is identical to the marginal likelihood given in (.). It is has been shown by cox that the method used to construct this likelihood gives max partial likelihood estimates that are consistent and a symptotic normally distributed with asymptotic covenians matrix estimated consistently by the inverse of the matrix of second partial derivatives of the long likelihood function.

Taking logarithms on both sides of eq (1.7.2)

$$\text{Log } L(\theta) = \sum_{i=1}^{k} \left[\underset{\sim}{x}_{(i)}\underset{\sim}{\beta} + \log \sum_{l \in R(t_{(i)})} e^{\underset{\sim}{x}_l \underset{\sim}{\beta}} \right]$$

It follows that the same asymptotic results for θ hold for estimation from partial likelihood as for the usual likelihood function. If the ties are present in the data the partial likelihood can be obtained by applying similar arrangement to the discrete logistic model. For this model the hazard relationship is given by

$$\frac{\lambda\ (t;\ \underset{\sim}{x})\ dt}{1 - \lambda\ (t;\ \underset{\sim}{z})\ dt} = \frac{\lambda_d(t)\ dt\ e^{\underset{\sim}{x}\underset{\sim}{\beta}}}{1-\lambda_d(t)\ dt} \qquad \text{......(1.7.3)}$$

where $\lambda_d(t)$ is an unspecified discrete hazard giving positive contributions at the observed failure times $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$.

A direct generalisation of the above argument can then be used to compute, at each failure time, the probability that the d_i failures should be those given the risk set and the multiplicity d_i.

A simple computation gives the conditional probability as the ith term in the product.

$$\prod_{i=1}^{k} \frac{e^{(S_i \beta)}}{\sum_{l \in R_{d_i}(t_{(i)})} e^{(S_i \beta)}} \qquad \text{......(1.7.4)}$$

where S_i = is the sum of the covariates associated with the d_i failures at $t_{(i)}$.

$$S_i = \sum_{j=1}^{d_i} x_{lj} \quad \text{and } l = (l_1, ..., l_{di})$$

$R_{di}(t_{(i)})$ is the set of all subsets of d_i items chosen from the risk set $R(t_{(i)})$ without replacement.

The partial likelihood (1.7.4) does not give rise to a consistent estimates of the parameter β in (1.7.1) if the ties arise by the grouping of continuous failure times. This inconsistency in the partial likelihood occurs since (1.7.4) must be thought of arising from the discrete model (1.7.3) and so estimates the odds ratio parameter β in that model since (1.7.3) does not arise as grouping of the continuous model, the two parameters do not have identical interpretations

1.7.3 BRESLOW'S MAXIMUM LIKELIHOOD METHOD :

The hazard function is approximated by a step function with discontinuities at each observed failure time i.e.

$$\lambda_0(t) = \lambda_i, \quad t_{(i-1)} < t \le t_{(i)} \qquad \text{......(1.7.5)}$$

$$(i = 1, 2, ..., k+1)$$

where $t_{(0)} = 0$ and $t_{(k+1)} = \infty$

If an individual censored in the interval $\left[t_{(i-1)}, t_{(i)}\right]$ is taken to have been censored at $t_{(i-1)}$.

The likelihood on the data is $L\left[\lambda_o(t_{(i)}); \underset{\sim}{\theta}\right] = \prod_{i=1}^{k} \lambda_o(t_{(i)})^{di}$

$$e^{S_i \underset{\sim}{\theta}} e^{-\int_o^{t_{(i)}} \lambda_o(u)du \sum_{l \in H(t_{(i)})} e^{\underset{\sim}{x}_i \underset{\sim}{\theta}}} \qquad \ldots\ldots(1.7.6)$$

where $H(t_{(i)})$ is the set of labels attached to the individuals either failing or censored at $t_{(i)}$. Using (1.7.5), (1.7.6) reduces to

$$L\left[\lambda_1, \lambda_2, \ldots \lambda_k, \underset{\sim}{\theta}\right] = \prod_{i=1}^{k} \left\{ \lambda_i^{di} e^{S_i \underset{\sim}{\theta}} e^{-\lambda_i (t_{(i)} \ldots t_{(i-1)}) \sum_{l \in R(t_{(i)})} e^{S_l \underset{\sim}{\theta}}} \right. \qquad \ldots\ldots(1.7.7)$$

$$\text{LogL} \left[\lambda_1, \lambda_2, \ldots, \lambda_k, \underset{\sim}{\theta}\right] = \sum_{i=1}^{k} d_i \log\lambda_i + S_i \underset{\sim}{\theta}$$

$$-\lambda_i (t_{(i)} \ldots t_{(i-1)}) \sum_{l \in R(t_{(i)})} e^{S_l \underset{\sim}{\theta}}$$

as the likelihood of $\lambda_1, \lambda_2, \ldots, \lambda_k$ and $\underset{\sim}{\theta}$ jointly no information is contained in the data about λ_{k+1} for any fixed $\underset{\sim}{\theta}$, (1.7.7) can be maximised with respect to λ_i. $(i = 1, 2, \ldots k)$ at

$$\hat{\lambda}_i = \frac{d_i}{\left[t_{(i)} - t_{(i-1)}\right] \sum_{l \in R(t_{(i)})} e^{x_i \underset{\sim}{\theta}}}$$

Substitution in (1.7.7) gives the maximum likelihood function of $\underset{\sim}{\beta}$ as proportional to (.).

The result (.) is obtained by this approach as the appropriate result when ties are present in the data where as the marginal or partial likelihood approach, (.) is obtained as an approximation to the exact result.

This approach can be criticised on the grounds that the data are in effect determining the model which is specified given the failure times to $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$.

It is also well known that maximizing a likelihood over a large number of muisance parameters can lead to misleading and biased results. Similar results can be obtained once the model (1.7.5) is selected by Baysian approach provided the distribution for log λ_i (i=1,2,...k) are taken to be independent uniform prices on $(-\infty,\infty)$ independently of the proper prior $p(\underset{\sim}{\beta})$.

1.8 : BAYSIAN INFERENCE :

We know that in parametric inference, the form of the population $f(x,\theta)$ is known while the parameter θ is unknown, however, we agree upon the parameter space i.e. the set of all possible values of the parameter which we denote by Ω. For example, in case of exponential density function

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta};\ x>0,\ \theta>0$$

The parameter space is $\Omega = \{ \theta | \theta > 0\}$.

In classical estimation theory the estimate of θ depends only on the sample values which we draw from $f(x;\theta)$ and as such the information about θ provided by the data only is taken into consideration. However, there may be situations in which we may wish to incorporate information about θ from other source as well.This additional information is called subjective judgment about the unknown parameter θ and can be combined with sample data using Baye's Theorem if expressible in the form of a probability distribution. There are cases in which θ can be regarded as a random variable with p.d.f $g(\theta)$. For example, in the case of exponential model, the mean life θ may be regarded as varying from batch to batch over time and this variation may be represented by a probability distribution over Ω.

Suppose that n items are placed on a test. It is assumed that their recorded lifetimes form a random sample say $X_1,X_2,\ldots\ldots,X_n$ which follow a distribution with p.d.f.$(x;\theta)$. To be specific we will assume to be real valued. Consider θ itself as a random variable with p.d.f. $g(\theta)$. Thus, the failure time p.d.f. $f(x;\theta)$ can be regraded as a conditional p.d.f. of x given θ i.e. $f(x|\theta)$. Where the marginal p.d.f. of θ is given by $g(\theta)$.

Therefore, the joint p.d.f. of $(X_1,X_2,\ldots,X_n,\theta)$ is expressed as

$$f\left[x_1,x_2,\ldots,x_n \mid \theta\right] = \prod_{i=1}^{n} f\left[x_i \mid \theta\right] g(\theta)$$

$$= L\left[x_1,x_2,\ldots,x_n \mid \theta\right] g(\theta) \quad \ldots(1.8.1)$$

The marginal p.d.f. of $(X_1,X_2,\ldots,X_n)$ is given by

$$P\left[X_1,X_2,\ldots,X_n\right] = P\left[X_1=x_1,X_2=x_2,\ldots X_n=x_n\right]$$

$$= \int_{\Omega} f(x_1,x_2,\ldots,x_n \mid \theta)\, d\theta \quad \ldots(1.8.2)$$

and the conditional p.d.f. of θ given data $(X_1,X_2,\ldots,X_n)$ is given by

$$\Pi\left[\theta \mid x_1,x_2,\ldots x_n\right] = \frac{f(x_1,x_2,\ldots,x_n \mid \theta)}{P\ (x_1,x_2,\ldots,x_n)}$$

$$= \frac{L(x_1,x_2,\ldots,x_n \mid \theta)\ g(\theta)}{\int_{\Omega} L(x_1,x_2,\ldots,x_n \mid \theta)\ g(\theta) d\theta} \quad ..(1.8.3)$$

Thus prior to obtaining the data $(X_1,X_2,\ldots X_n)$ the variations in θ were represented by $g(\theta)$ known as prior distribution of θ, however, after the data $(X_1,X_2,\ldots X_n)$ has been observed, in the light of new information the variations in λ are represented by $\Pi\left[\theta \mid x_1,x_2,\ldots x_n\right]$, the posterior distribution of θ. The uncertainty about the parameter θ prior to the experiment is represented by the prior p.d.f. $g(\theta)$ and the same after the experiment is represented by the posterior p.d.f. $\Pi\left[\theta \mid x_1,x_2,\ldots x_n\right]$ in (1.8.3).

In case the prior distribution of θ is discrete the integral sign in (1.8.3) is replaced by summation over Ω. This approach of development of posterior distribution is known as Baysian after River End Thomas Bayes an English Minister who lived in the 18th century. The process is a straight forward application of Bayes Theorem. Once the posterior distribution has been obtained, it becomes the main object of study.

Any statistical inference about θ may be draw with help of this posterior distribution.

Under the squared error loss function

$$L(\theta^*,\theta) = (\theta^*-\theta)^2 \qquad \text{......(1.8.4)}$$

Where θ^* is an estimate of θ.

The Bayes point estimate $\hat{\theta}$ of θ is defined as the posterior expectation of θ given the data $X = (X_1,X_2,...X_n)$ i.e.

$$\hat{\theta} = E\left[\theta \mid \underset{\sim}{X}\right] = \int_{\Omega} \theta\, \Pi\left[\theta \mid \underset{\sim}{X}\right] d\theta \qquad \text{......(1.8.5)}$$

It can be observed that for $\theta^* = \hat{\theta}$, the loss function defined above will attain its minimum value. A100$(1-\alpha)$% Bayes confidence interval (θ_1,θ_2) for θ may be obtained from

$$\int_{\theta_1}^{\theta_2} \Pi\left[\theta \mid \underset{\sim}{X}\right] d\theta = 1-\alpha \qquad \text{......(1.8.6)}$$

For testing hypothesis $H_0 : \theta \in H_0$ Vs $H_1 : \theta \in H_1$ where H_0 and H_1 are mutually exclusive sets of the parametric space H, we can make two decisions D_0 and D_1, D_0 means H_0 is true while D_1 implies H_1 is true. Now for wrong decision we have to anticipate some positive loss defined by

$L(\theta,D_o)$ = Loss incurred when decision D_o is made and θ is true value of the parameter.

$$= \begin{cases} 0 & \theta \in H_o \\ a(\theta) & ,\theta \in H_1 \end{cases}$$

$L(\theta,D_1)$ = Loss incurred when decision D_1 is made and θ is true value of the parameter.

$$= \begin{cases} b(\theta) & ,\theta \in H_o \\ 0 & ,\theta \in H_1 \end{cases}$$

Now in Bayes testing procedure we reject H_o if average loss for decision D_o is greater than average loss for decision D_1, i.e.

$$\text{if } \int_{H_1} a(\theta)\ \Pi\ (\theta|\underset{\sim}{x})\ d\theta > \int_{H_o} b(\theta)\ \Pi\ (\theta|x)\ d\theta \quad \ldots(1.8.7)$$

from equation (3), we have

$$\Pi\ (\theta|\underset{\sim}{x})\ \alpha\ \ L(x_1,x_2,\ldots,x_n|\ \theta)\ g(\theta)$$

which shows that for large samples the posterior is more dominated by the likelihood $L(x_1,x_2,\ldots,x_n|\theta)$ than the prior $g(\theta)$. Therefore as n tends to infinity or large, the Bayes estimates of θ will tend to its classical estimates showing thereby that in large samples the choice of the prior distribution is not very crucial.

1.9 : A BAYSIAN ANALYSIS OF THE PROPORTIONAL HAZARDS MODEL :

In this section we gives a brief outline of a non parametric Baysian analysis of the survival time data arising from the proportion a hazards model.

1.9.1 NON PARAMETRIC BAYSIAN PROCEDURES :

Here we consider specific application of some non-paramatric Baysian procedures to survival distributions. Suppose that the survivor function (conditional) of the random variable T is

$$P[\, T \geq t \mid \lambda_o] = e^{-\lambda_o(t)}$$

which is conditional on $\lambda_o(.)$. Since λ_o is the parameter in the model, is the realization of a stochastic process to be defined. Consider a partition of $[0,\infty)$ into a finite number k of disjoint intervals $[\theta_o = 0,\theta_1)$ $[\theta_1,\theta_2),\ldots,[\theta_{k-1},\theta_k=\infty)$ and defined the hazard contribution of ith interval as

$$\theta_i = P\left[T \in [\theta_{i-1},\theta_i) \mid T \geq \theta_{i-1},\lambda_o\right]$$

$$\text{if } P\left[T \geq \theta_{i-1} \mid \lambda_o\right] > 0 \qquad \ldots\ldots(1.9.1)$$

otherwise $\theta_i = 1$ $(i = 1,2,\ldots,k)$

$$\lambda_o(\theta_i) = \sum_{j=1}^{i} -\log\ (1-\theta_j) = \sum_{j=1}^{i} r_j \qquad \ldots\ldots(1.9.2)$$

where $r_j = -\log e\ (1-\theta_j)$

Doksum (1974) has considered this situation and has shown that a probability distribution can be specified on the space $\{\lambda_o(t)\}$ by specifying the finite dimensional distributions of $\theta_1,\theta_2,\ldots,\theta_k$ for each partition $[\theta_{i-1},\theta_i)$,

$i = 1,2,...,k$. Accordingly independent prior probability densities can be specified for $\theta_1,\theta_2,...,\theta_k$ subject to some consistency conditions and the resulting processes are called tailfree or neutral to the right by Doksum.

It is clear from (1.9.2) that $\{\lambda_o(t)\}$ is the non-decreasing independent increments process. This process in $\lambda_o(t)$ is called Subordinator by Iangman (1975).

CHAPTER-2

A DISCRETE SOFTWARE RELIABILITY GROWTH MODEL WITH LEADING AND DEPENDENT ERRORS

2.0 : INTRODUCTION :

Software reliability modelling is very important due to the fact that it is not possible to produce fault free software. The fault in the software occur due to human imperfection. These faults manifest themselves in terms of failure when the software is run. Testing phase in the software development process aims at detecting and removing these faults(errors) and making the software more reliable. Thus it is very important to evaluate software reliability during testing phase, based on software error data analysis. Modes concerned with the relationship between cumulative number of error detected through software testing and time span of testing are called software reliability growth models (SRGMs). Based on non-homogeneous Poisson Process (NHPP), several SRGHs [1,4-6,8-9] have been developed to predict remaining errors in the software and to evaluate measures such as mean time between failures, software reliability etc. Moreover, each software system is developed for a different objective and so it is not possible to develop an SRGM which can analyze failure data for all software systems.

Most of the SRGMs developed use calendar time or CPU time as the unit of software error detection/removal period. However, at times the number of test runs can be a more appropriate unit of software fault detection/removal period. Such as SRGM is called a discrete SRGM and relates the number of faults detected/removed to the number of test runs during the testing phase. A test run can be a single computer test run or a series of computer test runs executed in an hour, day, week or even month. Very few discrete SRGMs, have been developed in the literature.

Mostly, the error detection/removal phenomenon has been described by the exponential or s-shaped SRGMs. The s-shaped error removal phenomenon can be attributed to error depending [6] or to time lag between the failure due to an error and its subsequent removal [8], Bittanti [1] attributes s-shapedness to increased error removal during the later part of the testing phase. None of these models, however, describes the interface between independent leading errors and errors whose removal is dependent on these leading errors

In this chapter, we propose a new discrete SRGM assuming, the software contains two types of errors, leading and dependent. A leading error is defined as One that is immediate removed on its causing a failure. A dependent error is termed as One whose removal is delayed until the corresponding leading error is removed. The removal of a leading error helps in isolating the cause of failure of its corresponding dependent error. Applicability of the model has been shown by applying it to several software error data cited in [2].

Besides, modelling a software error detection process, it is also of utmost importance to know when to stop testing and release the software for use. Several criteria have been suggested in this regard [3-5, 7]. In this chapter, we also discuss a release policy for the proposed discrete SRGM by minimizing cost subject to discrete failure intensity not exceeding a specified value. We first estimate the parameters of the proposed SRGM by the method of maximum likelihood using software error data cited in [2]. Using the estimated values, we discuss the optimal release policy based on cost and intensity criteria. Results are illustrated by numerical examples.

2.1 : ASSUMPTIONS :

1. Software is subject to failures at random test runs caused by errors remaining in the software.
2. The error removal phenomenon is modelled by NHPP.
3. When a failure occurs, an intermediate effort is made to detect the error causing the failure and remove the error.
4. The errors in the software are divided into two categories : Leading (Independent) Faults Dependent Faults
5. The number of errors in the software is finite and is the sum of leading and dependent errors.
6. The expected discrete failure intensity for leading errors is proportional to the current remaining leading errors.
7. The expected discrete failure intensity for dependent errors is proportional to the current remaining dependent errors and the ratio of leading errors removed to the total number of errors.
8. The error removal process does not introduce any new errors in the software.
9. Software life cycle is assumed to be more than the optimal number of test cases/runs before releasing the software.

10. Software is never released without testing.

11. Corresponding to the error detection phenomenon at the manufacturer/user end, there exists an equivalent error detection phenomenon at the user/manufacturer end.

2.2 : NOTATIONS :

a : expected initial error content in the software, $a>0$.

$a_1(a_2)$: expected initial leading (dependent) error content in the software, $a_1>0$, $a_2\geq 0$, $a=a_1+a_2$.

$b(c)$: proportionality constant for leading (dependent) errors, $0<b<1$, $0\leq c<1$.

p : proportion of leading errors in the software, $0<p\leq 1$, $a_1=p.a$

N : Test run lag between removal of leading and dependent errors, $N\geq 0$.

$m_1(n)(m_2(n))$: mean value function for leading (dependent) errors, $m_1(n)=0$ for $n\leq 0$, $m_1(\infty)=a_1$, $m_2(n)=0$ for $n\leq N$, $m_2(\infty)=a_2$.

$m(n)$: $m_1(n) + m_2(n)$.

$\lambda(n)$: failure intensity for $m(n)(\lambda(0) = 0)$.

$C_1(C_2)$: Cost of fixing a leading error before (after) release of the software, $C_2>C_1>0$.

$C_3(C_4)$: Cost of fixing a dependent error before (after) release of the software, $C_4 > C_3 > 0$.

C_5 : Cost of a test run.

$C(n)$: Total expected software cost incurred during software life cycle, when the software is released after n test runs.

n_l : Software life cycle in terms of number of runs.

λ_o : desired failure intensity.

n^* : optimal number of test runs executed before releasing the software.

2.3 : MODEL ANALYSIS :

The number of leading errors removed on the $(n+1)^{th}$ test run as per assumption (6) may be expressed as

$$m_1(n+1) = m_1(n) = b(a_1 - m_1(n)) \qquad ......(2.1)$$

solving (2.1), under the initial condition $m_1(0) = 0$, we have

$$m_1(n) = a_1(1-(1-b)^n) \qquad ......(2.2)$$

equation (2.2) models the leading error removal phenomenon. The failure intensity for leading errors is given by

$$a_1 b(1-b)^n$$

it may be noted that failure intensity for leading errors decreases as n increases.

The number of independent errors removed on the $(n+1)^{th}$ test run, according to assumption (7) is expressed as

$$m_2(n+1) - m_2(n) = C(a_2 - m_2(n)) \frac{m_1(n+1-N)}{a} \qquad ......(2.3)$$

solving (2.3), under initial condition $m_2(0) = 0$, we get

$$m_2(n) = a_2\left[1 - \prod_{x=0}^{n-N} \left[1 - \frac{Ca_1}{a} (1-(1-b)^x)\right]\right] \qquad ...(2.4)$$

The failure intensity for dependent errors is given as

$$a_2 \prod_{0}^{n-N} \left[1 - \frac{Ca_1}{a} \left[1-(1-b)^x\right]\right] \frac{Ca_1}{a} \left[1-(1-b)^{n-N+1}\right]$$

It may be noted that failure intensity for dependent errors increases for all n (>N) satisfying

$$\frac{Ca_1}{a} \left[1-(1-b)^{n-N}\right] \left[1-(1-b)^{n-N+1}\right] < b(1-b)^{n-N}$$

and then decreases.

Now,

$$m(n) = m_1(n) + m_2(n)$$

$$= a - a_1(1-b)^n - a_2 \prod_{x=0}^{n-N} \left[1 - \frac{Ca_1}{a} \left[1-(1-b)^x\right]\right]$$

for simplicity, we assume test run lag to be negligible, i.e. N=0, so

$$m(n) = a - a_1(1-b)^n - a_2 \prod_{x=0}^{n} \left[1 - \frac{Ca_1}{a} \left[1-(1-b)^x\right]\right] \qquad ......(2.5)$$

$$= a\left[1 - p(1-b)^n - (1-p) \prod_{x=0}^{n} \left[1 - pc \left[1-(1-b)^x\right]\right]\right] \qquad ..(2.6)$$

equation (2.5) or (2.6) represents the expected number of errors removed in test runs.

The failure intensity for m(n) is

$$\lambda(n+1) = m(n+1) - m(n)$$

$$= a_1 b(1-b)^n + \frac{a_2 C a_1}{a}\left[1-(1-b)^{n+1}\right]$$

$$\prod_{x=o}^{n}\left[1 - \frac{Ca_1}{a}\left[1-(1-b)^x\right]\right] \quad ......(2.7)$$

It may be noted that either $\lambda(n)$ decreases for all $n>1$ or increases for $n \le n_x$ and decreases for $n > n_x$ where n_x (>1) satisfies

$$\lambda(n_x-1) < \lambda(n_x) > \lambda(n_x+1)$$

i.e., $\lambda(n)$ is maximum for $n = n_x$

2.4 : PARAMETER ESTIMATION :

The proposed mean value function m(n) has four unknown parameters. To estimate the parameters, we use the method of maximum likelihood.

Suppose, data is available for k observed pairs (n_i, y_i) $(i=1,2,...,k)$, where y_i is the cumulative number of faults removed by n_i test runs $(0 \le y_1 \le y_2 --- \le y_k)$ $(0<n_1<n_2---n_k)$. The likelihood function for the unknown parameters with m(n) in (2.6) is

$$L\left[a,b,c,p \mid (n_i,\ y_i)\right] =$$

$$\exp\left[-m(n_k)\right] \prod_{i=1}^{k} \frac{\left[m(n_i) - m(n_{i-1})\right]^{y_1 - y_{i-1}}}{(y_i - y_{i-1})\ !} \qquad \text{......(2.8)}$$

with $n_o = y_o = 0$. Taking Log of (8), we get

$$\ln\left[L(a,b,c,p \mid (n_i,y_i))\right] =$$

$$-m(n_k) + \sum_{i=1}^{k} (y_i - y_{i-1})\ \ln\left[m(n_i) - m(n_{i-1})\right]$$

$$-\sum_{i=1}^{k} (y_i - y_{i-1})\ ! \qquad \text{......(2.9)}$$

From (2.9), maximum likelihood estimates of a,b,c and p are obtained using DNCONF subroutine of IMSL MATH Library, under the following constraints

$$0 < a < \infty$$

$$0 < b < 1$$

$$0 \le c < 1$$

$$0 \le p \le 1$$

We have applied the proposed model for the following four real discrete software failure data sets cited in [2]

DS1 - The failure data is for a command, control and communication system software tested for twelve months. During this period 2657 errors were removed.

DS2 - The software failure data is for a command and control system software. The software was tested for fifteen weeks and 1138 errors were removed.

DS3 - The data is for a communication and control system software tested for fifteen weeks during which period 1483 errors were removed.

DS4 - This data set is also for a command and control system software tested for fifteen weeks and 2702 errors were removed.

The following table gives the maximum likelihood estimates of the proposed model parameters a,b,c and p for the four data sets described above :

	ESTIMATES OF			
DATA SETS	A	B	C	P
DS1	3115	0.1642	0.2104	0.7929
DS2	1385	0.1339	0.0743	0.8555
DS3	2562	0.4407	0.2533	0.1855
DS4	3942	0.0742	0.0	1.0

From the estimates of b, c and p obtained for the four failure data sets, it is observed that DS1 and DS2 have a high percentage of leading errors, while DS3 has a low percentage of leading errors and DS4 does not have any dependent errors and all the errors are leading. The model for DS4 data set reduces to a discrete exponential model. From the estimates obtained for the parameters, it is evident that the software either may not contain any dependent errors or it may contain a varying proportion of dependent errors. The above mentioned data sets simply justify the existence of such a discrete a software error removal phenomenon. Figures 1 to 4 show the graphs of actual and estimated failures for the above data sets.

2.5 : OPTIMAL SOFTWARE RELEASE PROBLEM :

It is very important for the software manager to know when to stop testing and release the software for use. Several researches have studied this problem based on different criteria for release [4-5 and reference cited there in]. In this chapter, we obtain the optimal number of test runs expected before release such that the total expected software cost during the life cycle of the software is minimized subject to discrete failure intensity not exceeding a prespecified value.

Mathematically, we may say

Minimize C(n) =

$$C_1 m_1(n) + C_2\left[m_1(n_l) - m_1(n)\right] + C_3 m_2(n) + C_4\left[m_2(n_l) - m_2(n)\right] + C_5 n \qquad ......(2.10)$$

subject to

$$\lambda(n) \leq \lambda_o \qquad \text{......(2.11)}$$

From (2.10),

$$C(n+1) - C(n) = -(C_2-C_1)\left[m_1(n+1) - m_1(n)\right] - (C_3-C_4)\left[m_2(n+1) - m_2(n)\right] + C_5$$

for simplicity, assuming $C_3 = C_1$ and $C_4 = C_2$, we have

$$C(n+1) - C(n) = -(C_2-C_1)\,\lambda(n+1) + C_5 \qquad \text{......(2.12)}$$

To study the behavior of the cost function, we consider the following cases

(i) When $\lambda(n)$ is decreasing for all n

if $C_5 > (C_2-C_1)\,\lambda(1)$, minimum cost is achieved for $n=0$, else

if $C_5 = (C_2-C_1)\,\lambda(1)$, minimum cost is achieved for $n=0$ and 1, else

if $C_5 < (C_2-C_1)\,\lambda(1)$, and there exists $n_1(>1)$ satisfying

$(C_2-C_1)\,\lambda(n_1-1) > C_5 > (C_2-C_1)\,\lambda(n_1)$, cost is minimum when $n = n_1-1$, else

if $C_5 < (C_2-C_1)\,\lambda(1)$ and there exists $n_1(>1)$ such that

$(C_2-C_1)\,\lambda(n_1-1) > C_5 = (C_2-C_1)\,\lambda(n_1)$

$C(n)$ is minimum for both $n=n_1$ and $n=n_1-1$

(ii) When $\lambda(n)$ increases for $n \leq n_x$ (>1) and decreases for $n>n_x$ such that $\lambda(n_x-1)<\lambda(n_x)>\lambda(n_x+1)$

if$C_5 \geq (C_2-C_1)\lambda(n_x)$, minimum cost is achieved for $n=0$,

else

if$C_5 < (C_2-C_1)\lambda(n_x)$, $C_5 \leq (C_2-C_1)\lambda(1)$ and there exists $n_1>n_x$ such that

$(C_2-C_1)\lambda(n_1-1) > C_5 > (C_2-C_1)\lambda(n_1)$, $C(n)$ is minimum for $n=n_1-1$, while if there exists $n_1>n_x$ such that

$(C_2-C_1)\lambda(n_1-1) > C_5 = (C_2-C_1)\lambda(n_1)$, $C(n)$ is minimum for both $n=n_1$ and $n=n_1-1$, else

if $C_5 < (C_2-C_1)\lambda(n_x)$, $C_5 > (C_2-C_1)\lambda(1)$ and there exist $n_2<n_x$ and $n_1>n_x$ such that

$(C_2-C_1)\lambda(n_2-1) < C_5 \leq (C_2-C_1)\lambda(n_2)$ and

$(C_2-C_1)\lambda(n_1-1) > C_5 > (C_2-C_1)\lambda(n_1)$,

$C(n)$ increases for $n\leq n_2-1$ and is minimum for $n = n_1-1$,

else if $(C_2-C_1)\lambda(n_2-1) < C_5 \leq (C_2-C_1)\lambda(n_2)$ and

$(C_2-C_1)\lambda(n_1-1) > C_5 = (C_2-C_1)\lambda(n_1)$,

$C(n)$ increases for $n\leq n_2-1$ and is minimum for both $n=n_1$ and $n=n_1-1$.

Now, for a specific intensity requirement λ_o(>0),

if $\lambda(n)$ is decreasing for all n and $\lambda(1) > \lambda_o$, these exists $n_i>1$ such that

$\lambda(n_i-1) > \lambda_o \geq \lambda(n_i)$, else if $\lambda(n)$ is increasing for $n \le n_x$ and decreasing for $n > n_x$ and $\lambda(n_x) > \lambda_o$, these may exist n_3 and n_4 $(0 < n_3 < n_x < n_4)$ such that

$$\lambda(n_3) \le \lambda_o, \ \lambda(n_3+1) > \lambda_o \qquad \text{......(2.13)}$$

$$\lambda(n_4) \le \lambda_o, \ \lambda(n_4-1) > \lambda_o \qquad \text{......(2.14)}$$

Combining the cost and intensity requirements, we may state the following theorem for optimal release policy (assuming unique n exists for minimum cost) theorem. Assume $C_2 > C_1 > 0$, $\lambda_o > 0$, $C_5 > 0$,

(a) $\lambda(n)$ is decreasing for all $n > 1$

(i) $\lambda(1) \le \lambda_o$

if $C_5 \ge (C_2 - C_1)\ \lambda(1)$, $n^* = 1$

if $C_5 < (C_2 - C_1)\ \lambda(1)$ and there exists n_1 such that $(C_2 - C_1)\ \lambda(n_1 - 1) > C_5 > (C_2 - C_1)\ \lambda(n_1)$, $n^* = n_1 - 1$

(ii) $\lambda(1) > \lambda_o$ and there exists $n_i > 1$ satisfying

$\lambda(n_i - 1) > \lambda_o \ge \lambda(n_i)$

if $C_5 \ge (C_2 - C_1)\ \lambda(1)$, $n^* = n_i$

if $C_5 < (C_2 - C_1)\ \lambda(1)$ and there exists n_1 such that $(C_2 - C_1)\ \lambda(n_1 - 1) > C_5 > (C_2 - C_1)\ \lambda(n_1)$,

$n^* = \max(n_i, n_1 - 1)$

(b) $\lambda(n)$ increases for $n \le n_x$ and decreases for $n > n_x$

(i) $C_5 \ge (C_2 - C_1)\ \lambda(n_x)$

if $\lambda(n_x) \le \lambda_o$, $n^* = 1$

if $\lambda(n_x)>\lambda_o$ and there exists n_3 satisfying (2.13),

$n^*=1$

if $\lambda(n_x) > \lambda_o$ and there exists n_4 satisfying (2.14),

$n^*=n_4$

(ii) $C_5<(C_2-C_1)\lambda(n_x)$, $C_5 \leq (C_2-C_1)\lambda(1)$ and there exists $n_1>n_x$ such that

$(C_2-C_1)\lambda(n_1-1) > C_5 > (C_2-C_1)\lambda(n_1)$

if $\lambda(n_x) \leq \lambda_o$, $n^*=n_1-1$

if $\lambda(n_x) > \lambda_o$ and there exist n_3 and n_4 satisfying (2.13) and (2.14) respectively, then

if $n_1 \geq n_4+1$, $n^*=n_1-1$, else

if $C(n_3) < C(n_4)$, $n^*=n_3$

if $C(n_3) > C(n_4)$, $n^*=n_4$

if $C(n_3) = C(n_4)$, $n^*=n_3$ or n_4

if $\lambda(n_x) > \lambda_o$ and there exists n_4 satisfying (2.14) then

if $n_1 \geq n_4+1$, $n^*=n_1-1$ else $n^*=n_4$

(iii) $C_5 < (C_2-C_1)\lambda(n_x)$, $C_5 > (C_2-C_1)\lambda(1)$ and there exists $n_2<n_x$ and $n_1>n_x$ satisfying

$(C_2-C_1)\lambda(n_2-1) < C_5 < (C_2-C_1)\lambda(n_2)$ and

$(C_2-C_1)\lambda(n_1-1) > C_5 > (C_2-C_1)\lambda(n_1)$

if $\lambda(n_x) \leq \lambda_o$

if $C(1) > C(n_1-1)$, $n^*=n_1-1$

if $C(1) = C(n_1-1)$, $n^*=1$ or n_1-1

if $C(1) < C(n_1-1)$, $n^*=1$

if $\lambda(n_x) \leq \lambda_o$ and there exist n_3 and n_4 satisfying (2.13) and (2.14) then

if $C(1) > C(n_1-1)$

if $n_1 \geq n_4+1$, $n^*=n_1-1$

if $n_3 \leq n_2-1$ or $(n_3 > n_2-1$ and $C(n_3) \geq C(1))$

if $C(1) < C(n_4)$, $n^*=1$

if $C(1) = C(n_4)$, $n^*=1$ or n_4

if $C(1) > C(n_4)$, $n^*=n_4$

if $C(n_3) < C(n_4)$, $n^*=n_3$

if $C(n_3) > C(n_4)$, $n^*=n_4$

if $C(n_3) = C(n_4)$, $n^*=n_3$ or n_4

if $C(1) = C(n_1-1)$

if $n_1 \geq n_4+1$, $n^*=n_1-1$ or 1, else $n^*=1$

if $C(1) < C(n_1-1)$, $n^*=1$

if $\lambda(n_x) > \lambda_o$ and there exists $n_4>n_x$ satisfying (2.14) then

if $n_1 \geq n_4+1$, $n^*=n_1-1$ else $n^*=n_4$

Other cases can be similarly discussed

2.6 : NUMERICAL EXAMPLES :

Using $\hat{a}=1385$, $\hat{b}=0.1339$, $\hat{c}=0.0743$, $\hat{p}=0.8555$, we discuss the optimal release policy for the software system described by DS2. We assume $C_1=5$, $C_2=10$, $C_5=40$, $n_l=150$ and $\lambda_0=.999$. Using these values, we have $n_1=27$ (C(n) is minimum for n=26) and $n_l=49$. Using *(ii)* of (a) in the theorem, we get $n^*=\max(n_l, n_1-1) = 49$. The intensity for n=49 is .995 and cost is 8948.4. If only cost was to be minimized, n^* would have been 26, cost as 8378.8 but intensity would have been quite high (8.0). Figures 5 and 6 show the graphs of cost and intensity functions respectively.

2.7 : CONCLUSION :

In this chapter, we have proposed a new discrete SRGM. At times discrete SRGMs are more suitable to describe software error detection/removal phenomenon than continuous SRGMs. In the proposed model the assumption of error independence has been relaxed.

Moreover, the proposed model can cater for various types of software growth modelling from pure exponential to highly s-shaped. Thus the proposed model can be applied to different testing environments.

Fig 1

DS 1

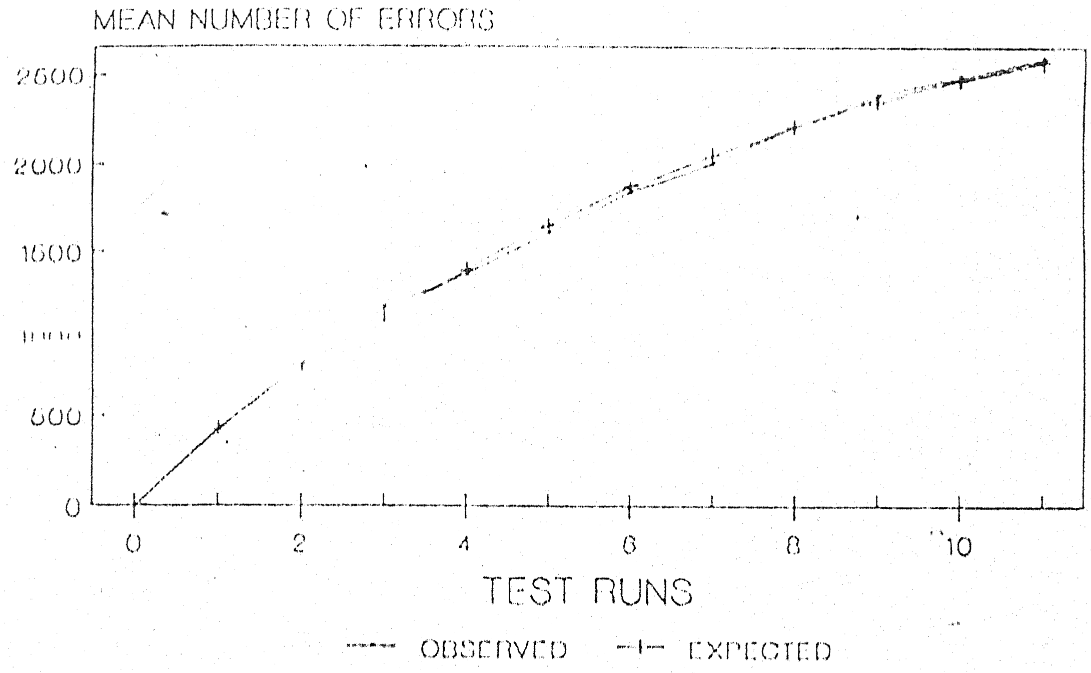


Fig 2

DS 2

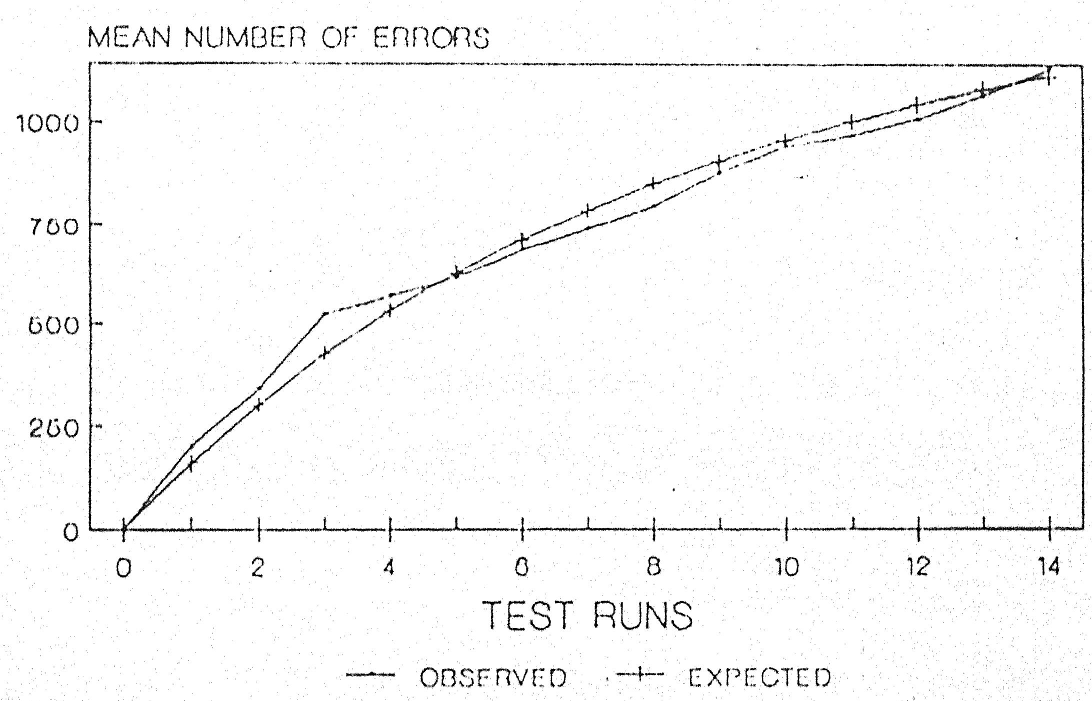


Fig 3
DS 3

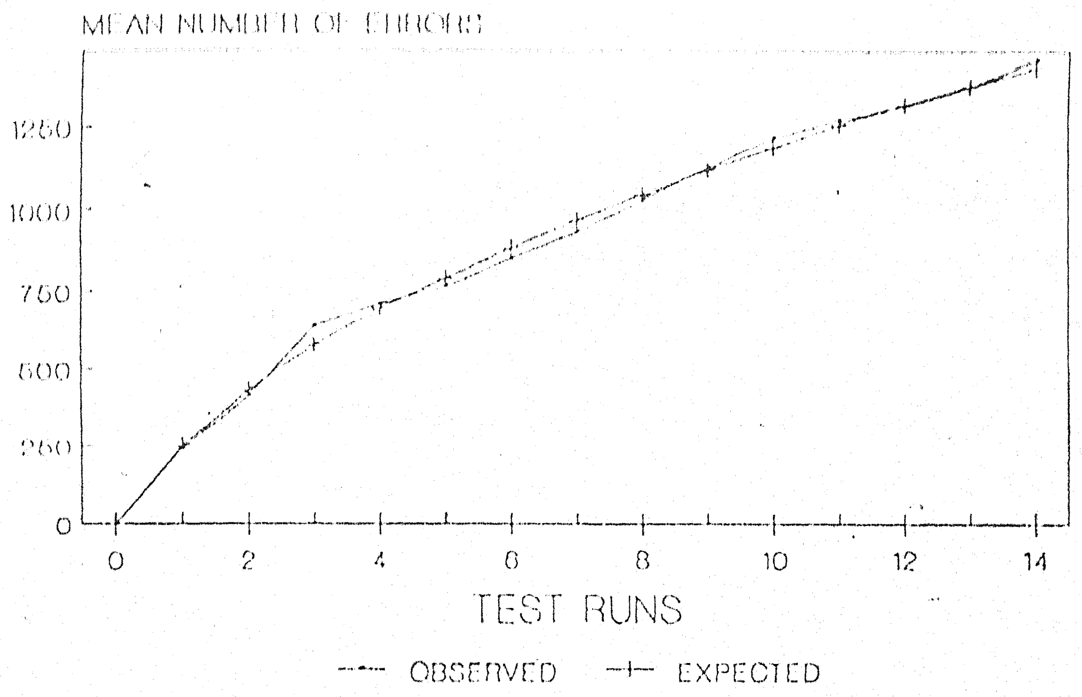


Fig 4
DS 4

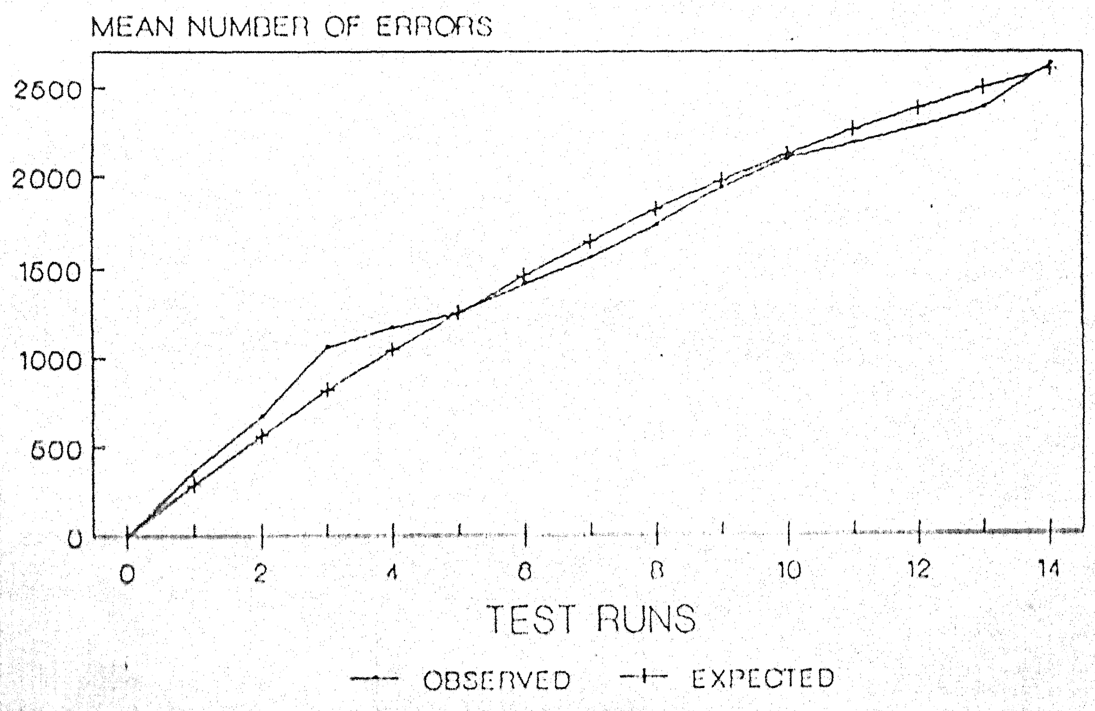


Fig 5

COST

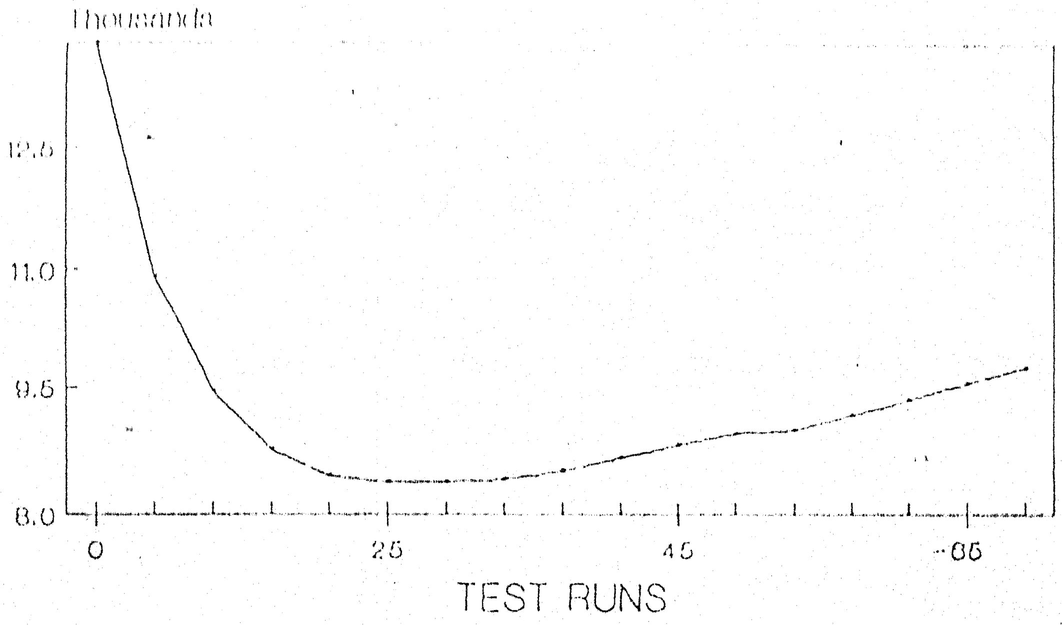
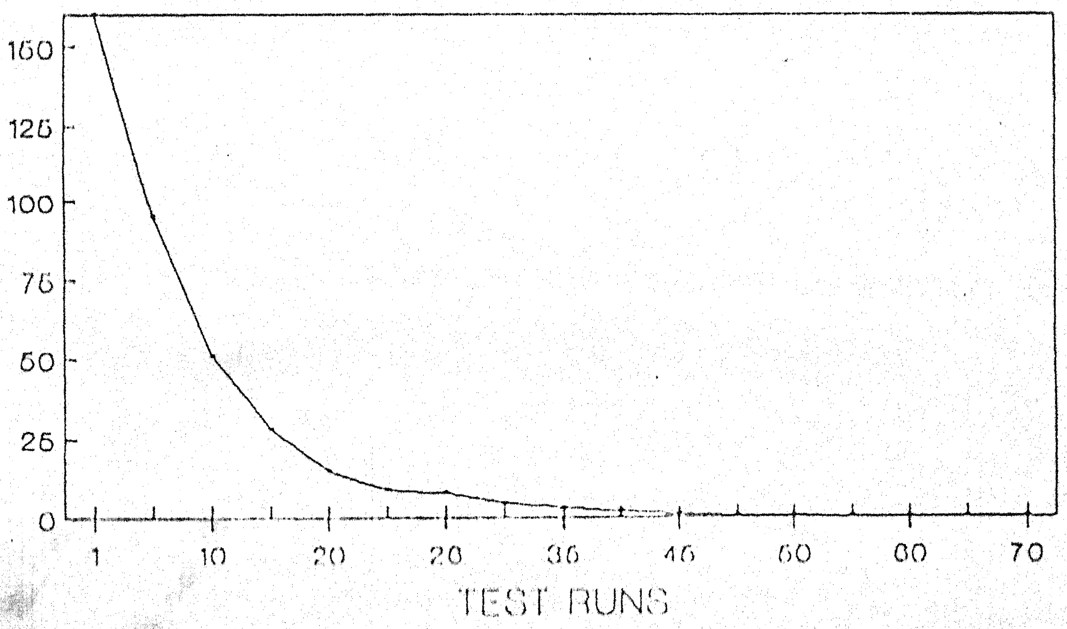


Fig 6

INTENSITY



CHAPTER-3

DISCRETE IMPERFECT DEBUGGING SOFTWARE RELIABILITY GROWTH MODEL

3.0 : INTRODUCTION :

In this chapter we propose a discrete software reliability growth model based on Non Homogeneous Poisson Process to describe the fault removal phenomenon under imperfect debugging environment. The learning process is taken into consideration by assuming that the probability of imperfect debugging phenomenon is independent on the faults remaining. The model has a flexible structure as it can describe different growth curves ranging from exponential to highly S-Shaped. The applicability of the model is shown by applying on data obtained from different software development projects.

Software Reliability Growth Models (SRGMs) are generally classified into two groups. The first contains the models which use the execution or calendar time as a unit of fault detection (Removal) period and such models are called continuous time models. The Second group contains models which use the test occasion (Cases) as a unit of fault detection (Removal) period and such models are called Discrete SRGMs. A test occasions (Case) can be a single computer test run or a series of computer test runs executed in an hour, day, week or even month. The test occasion (Case) includes the computer test run as well as length of time spent to visually inspect the software source code, Brooks and Motley [1], whereas a computer test run is a set of software input variables arranged in a certain manner to test the functional performance of a particular part of the software system. A large number of models have been developed in the first group while fewer are there in the second group Yamada and Osaki [8] proposed two discrete SRGM's, Kapur et al [6] proposed a general discrete software reliability growth model based on the assumptions that software may contain several type of faults. In all these models the fault removal process (Fault Debugging) is assumed to be perfect i.e. when an attempt is made to remove a fault, it is removed with certainty. This assumption may not be realistic. Due to the complexity of the software system and the incomplete understanding of the software requirements, Specifications and structure, the testing team may not be able to remove the fault perfectly and the original fault is replaced by another fault. The new fault may generate new failure when this part of the software system is transversed during the testing. The fault can be removed perfectly when the testing team properly understand the nature of the fault and takes the necessary steps to remove it. The multiple removal of the original fault and their successors i.e. the fault which replaces the original fault, slows downs the removal of the original fault and gives rise to S-Shaped Growth Curve. The concept of imperfect debugging was first introduced by Goel [2]. He introduced the probability of imperfect debugging in J.M. Model [4]. Kapur and Garg [5] introduced the imperfect debugging in Goel and Kumoto [3] Model. They assumed that the fault removal rate per remaining faults is reduced due to imperfect debugging. Thus the number of failures observed by time infinity is more than the initial fault content. Although these two models describe the imperfect debugging phenomenon yet the software reliability growth curve of these models is always exponential. Moreover, they assume that the probability of imperfect debugging is independent of the testing time. Thus they ignore the role of the learning process during the testing phase by not accounting for the experience gained with the progress of software testing. Actually, the probability of imperfect debugging is supposed to be maximum in the early stage of testing phase and is supposed to reduce with the progress of testing. Xia et al [7] also proposed an SRGM considering the role of learning process in the education of the probability of imperfect debugging. This model is based on sound assumptions but the determination of its parameters requires extra informations such as initial value of the probability of perfect debugging and the value of the learning factor. This information requires collection of extra data to use the model. Moreover, all these models are continuous time models.

In this chapter, we propose a discrete time SRGM based on Non-Homogeneous Poisson Process (NHPP) to describe the fault removal phenomenon under imperfect debugging environment. The learning process is taken into consideration by assuming that the probability of imperfect debugging is dependent on the number of faults remaining (Removed). The model has a flexible structure and can thus describe different growth curves. Further, the model is tested on a real software fault data obtained from various software development projects. The data sets are cited from Brooks and Motley [1].

3.1 : ASSUMPTIONS :

1. The fault removal/failure observation phenomenon follows Non-Homogeneous Poisson Process (NHPP).
2. The Software System is subject to failures at random times caused by software faults remaining in the software.
3. The expected number of failures observed between the n^{th} and $(n+1)^{th}$ test run occasion is proportional to the expected number of faults remaining in the software.
4. On the observation of a software failure, the efforts to remove the cause of the failure (the fault) may not be perfect and thus another version of the fault may replace the original fault.
5. The rate of imperfect debugging is decreasing with the testing time and is proportional to the number of faults remaining in the software.
6. The imperfect fault debugging does not increase the initial fault content.

3.2 : NOTATIONS :

a = The initial fault content in the beginning of the testing.

b = The Removal rate per remaining Fault.

c = The Initial imperfect debugging rate per fault.

$m_r(n)$ = The Expected Mean Number of Original faults removed by the n^{th} test occasion (Run).

3.3 : MODEL ANALYSIS AND FORMULATIONS :

Under assumptions (2,4), the expected number of original faults removed between the n^{th} and $(n+1)^{th}$ test runs satisfies the following difference equation :

$$m_r(n+1)-m_r(n) = b\left[a-m_r(n)\right]-c\left[\frac{a-m_r(n+1)}{a}\right]\left[a-m_r(n)\right] \quad ......(3.1)$$

The first term $b\left[a - m_r(n)\right]$ represent the intensity of faults debugged, while the negative term represents the intensity of imperfectly debugged faults. In other words, the intensity of faults removed is the intensity of faults debugged minus the expected intensity of the imperfectly debugged faults. To elaborate further, the initial imperfect debugging rate c is decreasing in the proportion of $\frac{a-m_r(n+1)}{a}$ as the testing progresses. Therefore, the remaining fault content $(a-m_r(n))$ is imperfectly debugged at the rate $c\left[\frac{a-m_r(n+1)}{a}\right]$.As the imperfectly debugged faults spawn new version of their own, consequently, these faults will generate more faults. Solving (3.1) using $m_r(0) = 0$, we get

$$m_r(n) = a\,\frac{(b-c)\;(1-(1-b)^n)}{(b-c) + c(1-b)^n} \qquad \text{Considering } \varphi = \frac{c}{(b-c)}$$

We get

$$m_r(n) = a\,\frac{c\,(1-(1-b)^n)}{1 + \varphi\,(1-b)^n} \qquad \text{......(3.2)}$$

From (3.2), we can see that $m_r(\infty) = a$. This indicates that the original faults are removed completely after a long time of testing.

3.4 : PARAMETER ESTIMATION :

The Maximum Likelihood Estimation (M.L.E) method is used to obtained the parameter estimates of the model given in (3.2). As the fault removal data used in this chapter are given in the form of pairs (n_i, x_i) $(i=1,2,...,k)$ where x_i is the cumulative number of faults removed by (n_i) test occasion $0<n<n_i<n_k$.

The likelihood function is given by

$$L\left[a,b,c\,|\,(n_i,x_i)\right] = \prod_{i=1}^{k} \frac{[m(n_i) - m(n_i-1)]^{x_i-x_{i-1}}}{(x_i - x_{i-1})!}\; e^{(-m(n_k))} \qquad \text{......(3.3)}$$

The parameter estimates are obtained by maximizing (3.3) with respect to each of the parameters. The DMCONE subroutine of IMSL MATH Library is used to maximize (3.3) and obtain the parameters estimates.

3.5 : DATA ANALYSIS :

To check the validity of the model, it is tested on two data sets cited from Brooke and Motley [1].

DS-1 :

The data is given in the form (n_i, x_i); $i=1,2,\ldots,12$ and the number of faults detected by the 12^{th} test occasion is 2657. The estimated values of the model parameters are

$$\hat{a} = 3169, \ \hat{b} = 0.148, \ \hat{c} = 0.0173$$

The proposed model estimates presence of imperfect debugging phenomenon in this project. This rate is much lower than the fault debugging rate per remaining fault (b). The fitting of the model is graphically illustrated in Fig. (1). It is clear that the model fits the fault the data excellently. It may be noted that the relationship between the cumulative number of faults and the test occasions is exponential.

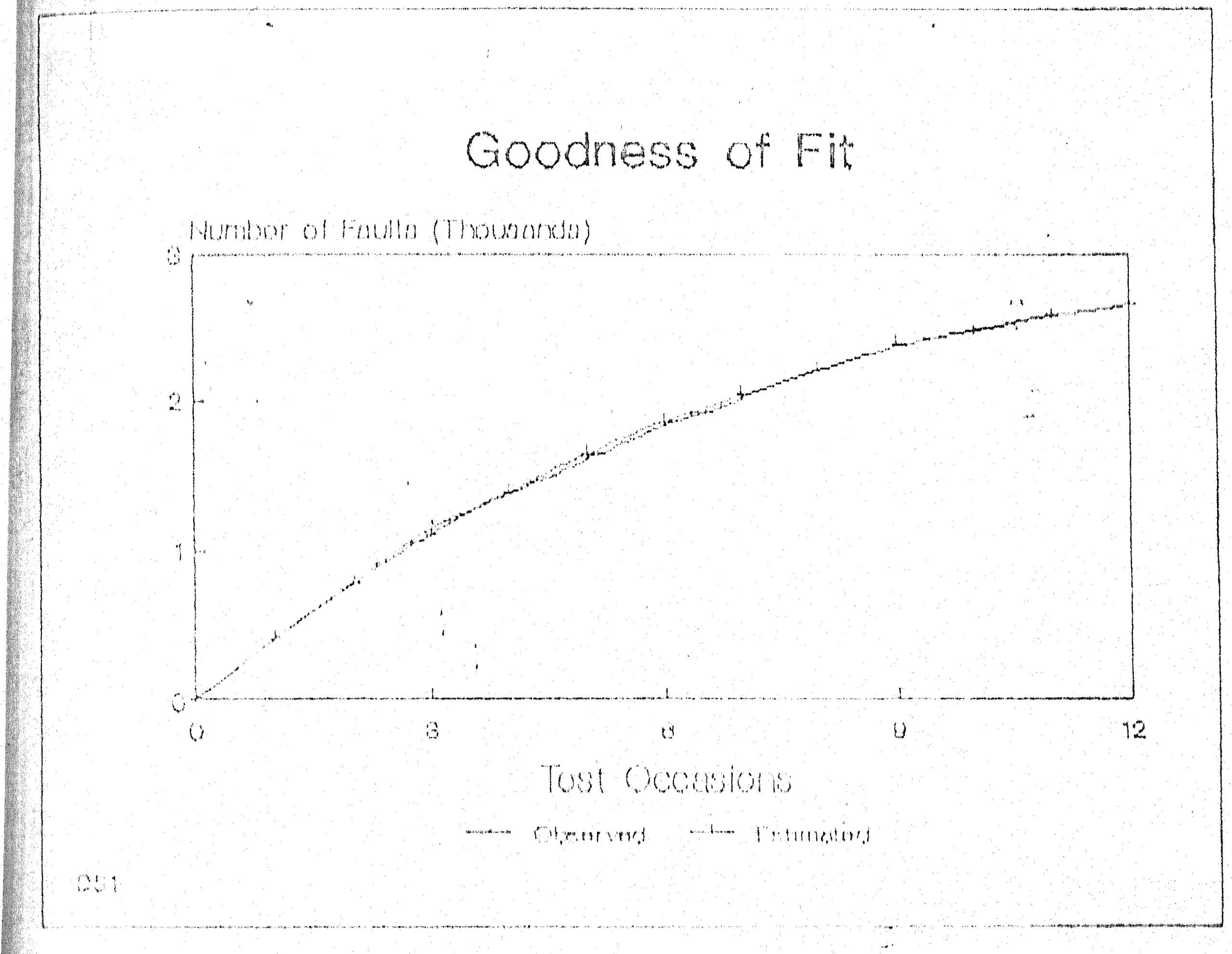
DS-2 :

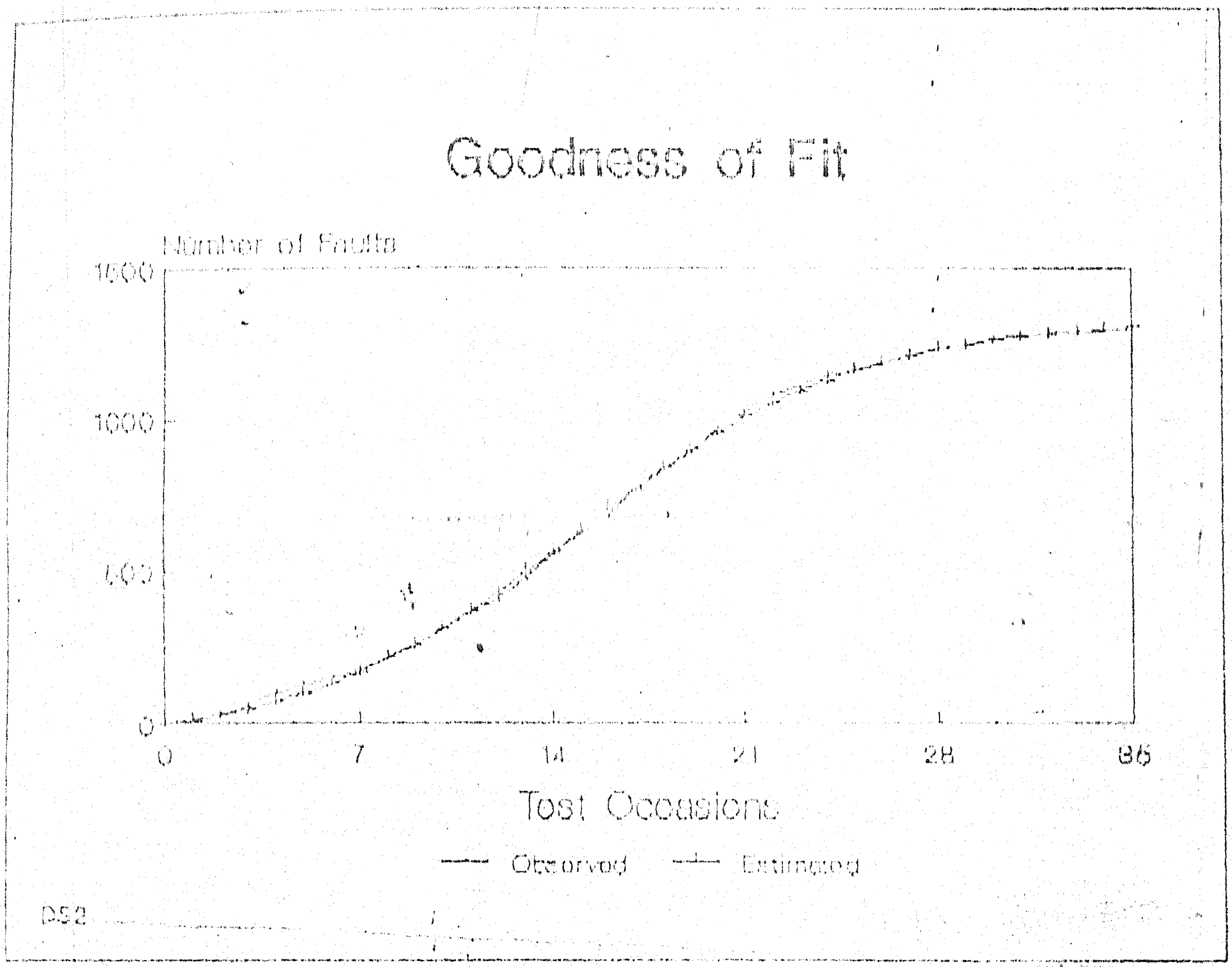
The software fault data is given in the form (n_i, x_i); $i=1,2,\ldots,35$, the number of faults detected by the 35^{th} test occasion is 1301. The estimated value of the model, parameters are :

$$a = 1325, \ b = 0.181, \ c = 0.172$$

The proposed model estimates the presence of imperfect debugging in this project. The initial value of imperfect debugging rate (c) is close to the value of (b).

This indicates that the imperfect debugging had a significant impact on the progress of the test at the early stages of the testing. This hypothesis is supported by fig. (2) as it is clearly seen that the relationship between the cumulative number of faults removed and the test occasions is not exponential (as the case in DS-1 Fig. (1)) but S-Shaped. In other words, the S-Shapedness in the reliability growth is attributed to the significant presence of imperfect debugging phenomenon. Further, fig. (2) graphically illustrates the goodness of fit of this model. It is clearly seen that the proposed model fits the observed fault data excellently.


Goodness of Fit
Number of Faults (Thousands)
3
2
1
0
0
3
6
9
12
Test Occasions
Observed
Estimated
051


Goodness of Fit
Number of Faults
1500
1000
500
0
0
7
14
21
28
35
Test Occasions
Observed
Estimated
D52

CHAPTER - 4

BAYSIAN ESTIMATION FOR THE GENERALISED THREE PARAMETER GAMMA DISTRIBUTION WITH TREE DIAMETER DATA

4.0 : INTRODUCTION :

Gamma distribution is also useful in many applications like reliability analysis, forestry and tree diameters. We have used the three parameter generalised Gamma density to model the distribution of tree diameters in forest stands. Gree E.J. et al (1994) considered Baysian estimation approach for three parameters Weibull density to model the distribution of tree diameters in forest stands. As we know, Weibull density is very much used in many applications such as reliability analysis, forestry and tree diameter data. But Gamma densities is also equally good for such applications. However, Krug, Nordheim and Giese (1984) has used Weibull density in modelling tree growth, survivorship and height distributions. Smith and Naylor (1987) find it difficult while estimating the parameters of Weibull density with reliability data using maximum likelihood estimation method. The estimate obtained for the location parameter is negative. Moreover, we will get negative estimate for location parameter m the case of modelling tree diameter distribution.

4.1 : MAXIMUM LIKELIHOOD ESTIMATION :

Let us consider generalised three parameter Gamma distribution as

$$f\left(x\,|\,\theta_1,\theta_2,\theta_3\right) = \frac{1}{\theta_2}\left(\frac{x-\theta_1}{\theta_2}\right)^{\theta_3-1} \frac{e^{-\left(\frac{x-\theta_1}{\theta_2}\right)}}{\Gamma(\theta_3)} \qquad \ldots\ldots(4.1)$$

$; \quad x \geq \theta_1$

$\theta_1, \theta_2 > 0$

where θ_1, is a location parameter

θ_2, is a scale parameter

θ_3, is a shape parameter

To obtain the maximum likelihood estimates of θ_1,θ_2 and θ_3, we differentiate the log-likelihood function with respect to θ_i (i=1,2,3). Let $X_1,X_2,\ldots,X_n$ be a random sample of size n drawn from the above distribution. The likelihood function is given by

$$L\left[\underset{\sim}{x}|\theta_1,\theta_2,\theta_3\right] = \prod_{i=1}^{n} f\left[x_i|\theta_1,\theta_2,\theta_3\right]$$

$$= \prod_{i=1}^{n} \frac{1}{\theta_2}\left[\frac{x_i-\theta_1}{\theta_2}\right]^{\theta_3-1} \frac{e^{-\left[\frac{x_i-\theta_1}{\theta_2}\right]}}{\Gamma(\theta_3)}$$

$$x_i \geq \theta_1$$

$$\theta_1,\theta_2,\theta_3 > 0$$

$$= \left[\frac{1}{\theta_2\Gamma(\theta_3)}\right]^{n} e^{-\sum_{i=1}^{n}\frac{(x_i-\theta_1)}{\theta_2}} \prod_{i=1}^{n}\left[\frac{x_i-\theta_1}{\theta_2}\right]^{\theta_3-1} \quad .....(4.2)$$

Taking logarithm on both sides of (4.2) the log-likelihood function is given by

$$\text{Log}_e L\ (\theta_1,\theta_2,\theta_3) = -n\ \log\theta_2 - \sum_{i=1}^{n}\frac{(x_i-\theta_1)}{\theta_2} - n\ \log\ \Gamma(\theta_3)$$

$$+ \sum_{i=1}^{n}(\theta_3-1)\ \left\{\log\ (x_i-\theta_1) - \log_e\theta_2\right\} \quad ....(4.3)$$

Differentiating with respect to θ_1, θ_2 and θ_3, we have

$$\frac{\partial\ \log_e L\ (\theta_1,\theta_2,\theta_3)}{\partial\theta_1} = \frac{n}{\theta_2} + \sum_{i=1}^{n}\frac{(\theta_3-1).(-1)}{x_i-\theta_1} = 0 \quad ......(4.4)$$

$$\frac{\partial\ \log_e L\ (\theta_1,\theta_2,\theta_3)}{\partial\theta_2} = -\frac{n}{\theta_2} + \sum_{i=1}^{n}\frac{(x_i-\theta_1)}{\theta_2^2} - \frac{n(\theta_3-1)}{\theta_2} = 0$$

$$-\frac{n\theta_3}{\theta_2} + \sum_{i=1}^{n}\frac{(x_i-\theta_1)}{\theta_2^2} = 0 \quad ......(4.5)$$

$$\frac{\partial\ \log_e L\ (\theta_1,\theta_2,\theta_3)}{\partial\theta_3} = \sum_{i=1}^{n}\log(x_i-\theta_1) - \sum_{i=1}^{n}\log\ \theta_2 - \sum_{i=1}^{\theta_3-1}\frac{n}{(\theta_3-1)} = 0$$

$$......(4.6)$$

From equation (4.4), we have

$$(\theta_3-1)\left\{\frac{1}{(x_1-\theta_1)}+\frac{1}{(x_2-\theta_1)}+\ldots+\frac{1}{(x_n-\theta_1)}\right\}=\frac{n}{\theta_2}$$

$$\sum_{i=1}^{n}\frac{1}{(x_i-\theta_1)}=\frac{n}{\theta_2(\theta_3-1)} \qquad ......(4.7)$$

From Equation (4.5), we have

$$-\frac{n\theta_3}{\theta_2}+\frac{n\bar{x}-n\theta_1}{\theta_2^2}=0$$

$$-n\theta_2\theta_3+n\bar{x}-n\theta_1=0$$

or

$$\boxed{\hat{\theta}_1=\bar{x}-\hat{\theta}_2\hat{\theta}_3} \qquad ......(4.8)$$

From Equation (4), we have

$$\log\prod_{i=1}^{n}(x_i-\theta_1)-n\log\theta_2=0$$

$$\log_e\theta_2=\frac{1}{n}\log\prod_{i=1}^{n}(x_i-\theta_1)$$

$$\boxed{\hat{\theta}_2=\left(\prod_{i=1}^{n}(x_i-\hat{\theta}_1)\right)^{1/n}} \qquad ......(4.9)$$

Since equation (4.7) and equation (4.9) does not provide any explicit solution for $\theta_{i'}$S (i'=1,2,3). Therefore, the maximum likelihood estimates can be obtained by iterative scheme.

4.2 : BAYESIAN MODEL :

Although Weibull distribution is generally used in a large number of diameter distribution but we have considered in this model generalised three parameter gamma distribution. In such types of problems investigator do not often reports the parameter estimates from individual samples. So data based priors are not readily available. Thus we choose Vague priors of θ_1, θ_2 and θ_3. As θ_2 and θ_3 are contained to be positive, we adopt Jeffreys (1961) prior for positive parameter θ_2 and θ_3.

i.e. $$g(\theta_2) \propto \frac{1}{\theta_2} \qquad \text{......(4.10)}$$

and $$g(\theta_3) \propto \frac{1}{\theta_3} \qquad \text{......(4.11)}$$

Regarding the parameter θ_1, we know that it is in the interval $\left[0, x_{(1)}\right]$. Where $x_{(1)}$ is the first order statistics i.e. the minimum diameter in the sample. Therefore, we choose the prior for the parameter θ_1 to be uniformly distributed R^{t}.

i.e. $$g(\theta_1) = \frac{1}{\theta_1}; 0 \leq \theta_1 < x_{(1)} \qquad \text{......(4.12)}$$

The condition $\theta_1 \leq x_{(1)}$ ensures that θ_1 will lie in the interval $\left[0, x_{(1)}\right]$. The baysian model is completed by specifying the prior distribution given in (4.10), (4.11) and (4.12).

The joint posterior distribution is then obtained by applications of Bayes Theorem

$$L\left[\underset{\sim}{x}\,|\,\theta_1,\theta_2,\theta_3;\underset{\sim}{x}\right] \;\alpha\; L\left[\underset{\sim}{x};\theta_1,\theta_2,\theta_3\right] \quad g\left[\theta_1,\theta_2,\theta_3\right]$$

where $L\left[\underset{\sim}{x};\theta_1,\theta_2,\theta_3\right]$ is the likelihood function and $g\left[\theta_1,\theta_2,\theta_3\right]$ is joint prior for θ_1,θ_2 and θ_3. Further, we will take the priors for θ_1,θ_2 and $\theta_3$to be independent where as dependencies in the posterior arises due to the sample data.

The full Baysian model is given by

$$L\left[\theta_1,\theta_2,\theta_3;\underset{\sim}{x}\right] \;\alpha\; \left[\frac{1}{\theta_2}\right]^n e^{-\frac{1}{\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)} \prod_{i=1}^{n}\left[\frac{x_i-\theta_1}{\theta_2}\right]^{\theta_3-1} \frac{1}{\theta_2}\cdot\frac{1}{\theta_2}$$

$$= \frac{1}{\theta_2^{n+1}\;\theta_3}\, e^{-\frac{1}{\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)} \prod_{i=1}^{n}\left[\frac{x_i-\theta_1}{\theta_2}\right]^{\theta_3-1} \quad \ldots\ldots(4.13)$$

$$x_{(1)} \geq \theta_1 > 0$$

$$\theta_2 > 0,\ \theta_3 > 0$$

The normalising estimate of the parameters for equation (4.13) are difficult to obtain analytical. However, we can apply stochastic simulation procedure to iterative scheme to estimate the parameter involved in equation (4.13). For instance, suppose that θ_{10},θ_{20} and θ_{30} are our initial values of the parameters θ_1, θ_2 and θ_3. We can also use Gibbs sampler which is based on iterative scheme. To make use of Gibbs sampler we must be able to sample from the full conditional distribution for each parameter. The full conditional for θ_1, θ_2 and θ_3 are given by

$$\prod\left(\theta_2|\theta_1,\theta_3;\underset{\sim}{x}\right) \propto \theta_2^{-(n\theta_3+1)}\, e^{-\sum_{i=1}^{n}(\frac{x_i-\theta_1}{\theta_2})} \quad ......(4.14)$$

$$\prod\left(\theta_3|\theta_1,\theta_2;\underset{\sim}{x}\right) \propto \theta_3^{-1}\prod_{i=1}^{n}\left(\frac{x_i-\theta_1}{\theta_2}\right)^{\theta_3-1} e^{-\sum_{i=1}^{n}(\frac{x_i-\theta_1}{\theta_2})} \quad ...(4.15)$$

$$\prod\left(\theta_1|\theta_2,\theta_3;\underset{\sim}{x}\right) \propto \prod_{i=1}^{n}\left(x_i-\theta_1\right)^{\theta_3-1} e^{-\sum_{i=1}^{n}(\frac{x_i-\theta_1}{\theta_2})} \quad ......(4.16)$$

Random numbers can be generated from $\prod\left(\theta_2|\theta_1,\theta_3;\underset{\sim}{x}\right)$ by drawing a random variable Z from Gamma distribution. Following an initial guess at the value of each parameter, the Gibbs sampler proceeds by iteratively generating a new value for each parameter. If θ_{10}, θ_{20} and θ_{30} are the initial values. We generate θ_{11} from $\prod\left(\theta_1|\theta_{10},\theta_{20};\underset{\sim}{x}\right)$ then θ_{21} from $\prod\left(\theta_2|\theta_{11},\theta_{20};x\right)$ and finally θ_{31} from $\prod\left(\theta_3|\theta_{11},\theta_{21};\underset{\sim}{x}\right)$. This constitutes one iteration of the sampler. In the next iteration, initial values are replaced by those from the first iteration and values from the first iteration by those from the second iteration and so on.

4.3 : CONCLUSION :

The main purpose here is to demonstrate the Baysian procedure for three parameter Gamma distribution which is better than any typical method found in literature. The numerical illustration can be done if the data of tree diameter distribution is available. It can be seen that the maximum likelihood estimation is problematic with tree diameter distribution because of negative estimates of location parameter θ_1. Baysian model is easy to fit and results are also according to our prior expectations. In particular, widely used adhoc rules such as Set $\hat{\theta}_1 = 0$ whenever $\hat{\theta}_1 < 0$ are necessary with Baysian inference. If the Gamma density is to be used in a tree diameter distribution then prediction would be probably be made using the model values of the joint posterior sample of the parameters, whereas in Baysian approach, once the samples have been obtained from the full posterior, these can be used to great advantage for simulating any predictive distribution of interest.

CHAPTER-5

A BAYSIAN ESTIMATION APPROACH TO PROPORTIONAL HAZARD MODELS FOR COVARIATES AND INSTITUTIONAL EFFECTS

5.0 : INTRODUCTION :

Institutional variation is an important factor to examine in a randomized clinical trials. in randomized clinical trials for comparing treatments for disease such as cancer, it is sometimes necessary to include patients from different institutions to compare the sample size in a reasonable period of time. One of the reasons to examine institutional variations is that the objective of clinical trial is to try to draw conclusion about the overall effect of therapy in the population. Since the institutions are not selected at random and only a small random subset or the patients are entered on trials with substitutional institutional variation it would not be clear exactly what effect would be seen in the general population. Another reason to examine institutional variation is that it might be possible to learn more about how the therapy should be given or to whom it should be given. There are several papers which deals with affect of institutions in clinical trials. Skene and Wakefield (1990) use the same type of hierarchical Baysian Structure to model data from multi center binary response trials. Boos and Brownie (1992) proposed rank based methods for analysing data from multi center trials with continuous or order categorical outcomes with a linear model structure.

Stangl and Greenhouse (1992) developed hierarchical Baysian survival model for examining institutional differences. Grey (1994) considered a Baysian analysis of institutional effects in a multi center cancer clinical trial. The structure of their model is quite different from the proportional hazard model used here.

5.1 : BAYSIAN MODEL :

A proportional hazard model is assumed for institutional effects and covariates. let x_{ijk} be covariate k for subject j from institution i.

Let us assume that there are N institutions with n_i cases/patients from institution i and (p-1) covariates with x_{ijp} as treatment variables.

Let $0 < t_o < t_1 <....< t_m$ be the boundaries of time intervals and set $I_l(t) = I(t_{l-1} < t < t_l)$.

The full hazard model for subject ij can be written as

$$h(t|x_{ij},\alpha,\beta,\theta_i)=e^{\sum_{l=1}^{m}\alpha_l I_l(t)+\theta_{io}+\sum_{k=1}^{p}x_{ijk}\beta_k+\theta_{i1}x_{ijp}} \quad ..(5.1)$$

$$j = 1,2,.....,n_i$$
$$;i = 1,2,.....,N$$
$$k = 1,2,.....,p-1$$

$$\log h(t|x_{ij},\alpha,\beta,\theta_i)=\sum_{l=1}^{m}\alpha_l I_l(t)+\theta_{io}+\sum_{k=1}^{p}x_{ijk}\beta_k+\theta_{11}x_{ijp} \quad ..(5.2)$$

where $\underset{\sim}{\alpha} = \begin{bmatrix}\alpha_1\\ \alpha_2\\ :\\ \alpha_m\end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\\ :\\ \beta_p\end{bmatrix}, \quad \theta_i = \begin{bmatrix}\theta_{io}\\ \theta_{i1}\end{bmatrix}$

are unknown parameters.

$$x_{ij} = \begin{bmatrix}x_{ij1}\\ x_{ij2}\\ :\\ x_{ijp}\end{bmatrix}$$

θ_{io} are the institutional deviations from an overall underlying log hazards $\sum_{l=1}^{m}\alpha_l I_l(t)$.

θ_{i1} is the deviation in the i^{th} institution from overall effect β_p.

Let us consider four covariates x_1,x_2,x_3 and x_4.

where x_1 : performance status

x_2 : months from diagnosis

x_3 : age in years

x_4 : prior therapy

$$\log h(t|x_{ij},\alpha,\beta,\theta_i)=\sum_{l=1}^{m}\alpha_l I_l(t)+\theta_{i0}+x_{ij1}\beta_1+x_{ij2}\beta_2$$

$$+x_{ij3}\beta_3+x_{ij4}\beta_4+\theta_{i1}x_{ij4} \quad ......(5.3)$$

Let us assume the following prior for the model given above

β_i follows double exponential distribution or Laplace distribution with probability density function

$$g(\beta_i) = \frac{1}{2}\, e^{-|\beta_i|}; \qquad -\infty<\beta<\infty$$

Let θ_i is independent and identically multivariate normally distributed with mean zero and variance unity.

$$\theta_i \overset{iid}{\sim} N(O,I)$$

$$g(\theta_i) = \frac{1}{\sqrt{2\pi}}\, e^{-\theta_i^2/2}$$

$$\alpha_l-\alpha_{l-1}|v \sim N(O,v^{-1}) \quad l = 1,2,...,m$$

and $\qquad v \;\Lambda\; \gamma\;(\mu,\lambda)$

i.e. $\qquad g(v) = \frac{r^{\mu}}{\Gamma(\mu)}\, e^{-\lambda v}\, v^{\mu-1}; \quad v\geq 0$

The prior for α restricts the magnitude of the jump between adjacent intervals in a piece wise constant model.

$$\text{Let } \delta_{ijl} = \begin{cases} 1 & \text{if subject (i,j) is observed to fail in time interval l} \\ 0 & \text{otherwise} \end{cases}$$

Let T_{ijl}=0 if the failure or censoring time < t_{l-1}. The likelihood function corresponding to this model is given by

$$L(\alpha,\beta,\theta_i) = \prod_{i=1}^{N} L_i(\alpha,\beta,\theta_i)$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{n_i} h(t|x_{ij},\alpha,\beta,\theta_i)\ e^{-\int_{o}^{T_{ijl}} h(t|x_{ij},\alpha,\beta,\theta)du}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{n_i} e^{\sum_{l=1}^{m} \alpha_l I_l(t)+\theta_{io}+x_{ij1}\beta_1+x_{ij2}\beta_2+x_{ij3}\beta_3+(\beta u+\theta_{i1})x_{iju}}\ e^{-T_{ijl}h(t|x_{ij}\alpha,\beta,\theta)}$$

$$= \prod_{i=1}^{n} e^{\sum_{j=u=1}^{n_1}\sum^{m}\delta_{ijl}\ \eta_{ijl}}\ e^{-T_{ijl}\ e^{\eta_{ijl}}}$$

$$= \prod_{i=1}^{N} \exp\left[\delta_{ijl}\ \eta_{ijl} - T_{ijl}\ e^{\eta_{ijl}}\right] \qquad \text{......(5.4)}$$

where $\eta_{ijl}=\alpha_l+\theta_{io}+x_{ij1}\beta_1+x_{ij2}\beta_2+x_{ij3}\beta_3+(\beta_4+\theta_1)x_{ij4}$ ...(5.5)

The joint posterior is then proportional to

$$= \left[\prod_{i=1}^{N} L_i(\alpha,\beta,\theta_i)\ g(\theta_i)\right] g(\alpha|v)\ g(v)\ g(\beta) \qquad \text{......(5.6)}$$

where g(.) function are is the prior densities. This is not easy to compute directly but it is possible to generate samples from the joint posterior using Gibbs sampling. In

Gibbs sampling observations are generated from the j joint posterior distribution by sampling from full conditional distributions. See Gelfand and Smith (1990), Gelfand et al (1990), Zejer and Karim (1991) and Clayton (1991).

The parameters that are require to specify the prior densities are λ and μ in the prior for v. Proper priors are used but the parameters are chosen to keep the priors fairly weak. For institutional effects this justified since there is little prior information on the magnitude of these parameters for the prior v, it is noted that 1/v is the variance of jumps in the log-underlying hazards at the boundaries of the time intervals. The question of the magnitude of survival difference can also be addressed using predictive distributions. The survival curve for the predictive distribution for a new case from institution i is given by the integral of

$$\delta(t) = e^{-\int_0^t h(x|x_{ij},\alpha,\beta,\theta_i)du}$$

over the posterior distribution where h(.) s given by (5.1) with Gibbs sampling, the integral over the posterior is calculated by averaging over the generated parameter value.

Data analysis has not been done because of non-availability of data.

5.2 : CONCLUSION :

This chapter deals with a Baysian estimation procedure for study the amount of institutional variation in a multicentre clinical trial using proportional hazard models. A hierarchical structure is used with prior for covariates coefficients as double exponential distribution and prior for institutional deviations θ_{io} as standard multivariate normal density with mean vector zero and variance-covariance matrix I. The prior for α restricts the magnitude of the jump between adjacent intervals $\alpha_l-\alpha_{l-1}$ is i:i:d normal variate with variance 1/v. Further the prior for v is a Gamma distribution with parameter μ and λ. The posterior distribution calculated using Gibbs sampling. The methods can not be applied to data from Lung Cancer trial because non-availability of data. This study can be proceeded further by applying it to Lung Cancer trial data. We can predict that there appears substantial variation in the treatment effect across institutions. Although the reason for this have not been identified. It would be possible to investigate this further through a detailed examination of the data from the institutions with extreme effects.

BIBLIOGRAPHY

[1] Bailey R.L. and Dell T.R. (1973) : *Quantifying Diameter Distributions with the Weibull Function* Forest Science 19, 97-104.

[2] Bittanti S, Bolzern, P. Pedrotti, E, Pozzi, M and Scattolini, R(1988) :*A Flexible Modelling approach for software reliability growth*, In software reliability modelling and identification, ed. S. Bittanti, Springer-Verlog, 1988, Berlin.

[3] Brooks W.D. and Motley R.W.(1980), *Analysis of discrete software reliability models*, Technical Report, RADC-TR-80-84, Rome Air Development Centre, 1980, New York.

[4] Brooke S.W. and Motley R.W. (1980) : *Analysis of Discrete Software Reliability Model*, Technical Report, RADC-TR-84, Rome Air Development Centre, New York.

[5] Boose D.D. and Brownie C. (1992) : *A Rank Based Mixed Model Approach to Multisite Clinical Trials*, Biometrics 48, 61-72.

[6] Caspi P.A. and Kouka E.F.(1984) : *Stopping Rules for a Debugging Process based on different Software Reliability Models*, Proc. int. conf. on fault tolerant computing, 1984, pp. 114-119.

[7] Clayton D.G. (1991) : *A Monte Carlo Method for Baysian Inference in Normal Data Models*, J.A.S.A., 85, 972-985.

[8] Crowder M.J., Kimber A.C., Smith, R.L. and Sweeting T.J. (1991) :*Statistical Analysis of Reliability Data*. London : Chapman & Hall.

[9] Dagpunar J. (1988) : *Principles of Random Variable Generation*, London : Oxford University Press.

[10] EK A.R., Issos, J.N. and Baily R.L. (1975) :*Solont for Weibull Diameter Distribution Parameters to Obtain Specified Mean Diameters*, Forest Science, 21, 290-292.

[11] Gelfand A.E., Hills S.E., Racine-Poon A. and Smith AFM (1990) : *Illustration of Baysian Inference in Normal Data Models using Gibbs Sampling*, JASA, 85, 972-985.

[12] Gelfand A.E. and Smith AFM (1990) : *Sampling Based Approaches to Calculating Marginal Densities*, J.A.S.A., 85, 398-409.

[13] Goel A.L. and Okumoto K.(1979) : *Time Dependent Error Detection Rate Model for Software Reliability and other Performance Measures*, IEEE Trans. Rel. R-28, 1979, pp. 206-211.

[14] Goel, A.L. (1980) : *Software Error Detection Model with Application*, Journ. Sys. Software, 1, 243-249.

[15] Goel, A.L.(1985) : *Software Reliability Models; Assumptions, Limitations and Applicability*, IEEE Trans Software Engg,SE-11, 1411-1423.

[16] Green E.J. and Straw Derman (1992) : *A Comparision of Hierarchical Bayes and Empirical Bayes Methods*, Forest Science, 38, 350-366.

[17] Grey Robert J. (1994) : *A Baysian Analysis of Institutional Effects on a Multicentre Clinical Trial*, Biometrics, Vol. 50, 244-253.

[18] Jeffreys H. (1961) : *Theory of Probability*, 3rd edition, London, Oxford University Press.

[19] Jelinski Z. and Moranda P.B., (1972) : *Software Reliability Research. In Statistical Computer Performance Evaluation ED. W. Freiberger*, Academic Press, New York, 465-484.

[20] Kapur P.K. and Garg R.B. (1990)(a) :*A Discrete S-Shaped Software Reliability Growth Model*, Journal of Sys Software.

[21] Kapur P.K. and Garg R.B. (1990)(b) : *Optimal Software Release Policies for Software Reliability Growth Model under Imperfect Debugging*, R.A.I.R.O.,24,295-305.

[22] Kapur P.K. and Garg R.B. (1989) : *A General Theortical Frame Work for Software Release Policies*, IEEE Trans Reliability.

[23] Kapur P.K. and Garg R.B.(1992) : *A Software Reliability Growth Model for an Error Removal Phenomenon*, Software Eng. Journal, vol. , No.-4, 1992, pp. 291-294.

[24] Kapur P.K., Yunes S. and Aggarwala S.(1994) : *Generalised Erlang Software Reliability Growth Models* - to be published in ASOR Bulletin.

[25] Krug A.G., Nordheim E. V. and Giese R.L. (1984) : *Determining Initial Values*.

[26] Musa J.D., Iannino, A. and OKumoto, K. (1984) : *Software Reliability*, McGraw Hill Book Co., New York.

[27] Musa J.D. (1975) : *A Theory of Software Reliability and its Application*, IEEE Trans. Software Engg., 312-327.

[28] Musa J.D. and Okumoto K. (1989) : *A Logarithm Poisson Execution Time Model for Software Reliability Measurement*, Proc 7th Inf. Conf. on Software Engg., Ortando, Florida, March 26-29, 230-238.

[29] Ohba, M. (1984) : *Software Reliability Analysis Models*, I.B.M. Journ. Rev. Dev. 28(4), 428-443.

[30] Ross, S.M. (1985) : *Software Reliability : The Stopping Rule Problem*, IEEE Trans. Software Engg., SE-11, 1472-1476.

[31] Skene A.M. and Wakefield J.C. (1990) : *Hierarchical Models for Multicentre Binary Response Statistics in Medicine*, 9, 919-929.

[32] Smith R.L. and Naylor J.C. (1987) : *A Comparision of maximum likelihood and Baysian Estimation for the Three Parameter Weibull Distribution*, Applied Statistics, 43, 201-211.

[33] Vallee F.M. and Rasot A. (1991) : *Evaluation Using NHPP Models*, Proc. Int. Symp. on Software Reliability Engineering May 18, 19 Austin, Texas, 157-167.

[34] Xia G., (1992) : *Mathematical Analysis of Software Reliability Growth Model*, Master Thesis, Deptt. of Mathematics, Royal Melbourne Institute of Technology (RMIT), Melbourne, Australia.

[35] Yamada S. and Osaki S. (1985) : *Discrete Software Reliability Growth Model*, App. Stochastic Models & Data Analysis 1, 65-77.

[36] Yamada S., Ohba M. and Osaki S. (1984) : *S-shaped Software Reliability Growth Models and their Applications*, IEEE Trans. Reliab., R-33, 289-292.

[37] Yamada S., Ohtera H. and Narihisa H. (1986) : *Software Reliability Growth Models with Testing Efforts*, IEEE Trans. Reliab., R-35(1), 19-23.

[38] Yamada S., Osaki S. and Narihisa H. (1985) : *A Software Reliability Growth Model with Two Types of Errors*,